医学統計学演習：資料 4

芳賀昭弘 *

1　復習：グラフ（ヒストグラム）の作成

R を起動し、次のように入力みよう。

```
> saikorome <- c(1:6)
> saikorodata <- rep(1/6, 6)
> saikoro <- data.frame(saikorome,saikorodata)
> hist(saikorome,breaks=seq(0.5,6.5,1),probability=T,ylim=c(0,1))
```

ヒストグラムがプロットされましたね？ 1 から 6 までの整数の値が確率 1/6 で起こる確率分布をヒストグラムで書くとすると図のようになります。本日のテーマは「確率分布」です。

2　確率分布

確率分布は次の条件を満たす必要があります。

- 生起確率は負にならない。
- 取り得る確率変数の生起確率を全て足す（連続変数では積分する）と 1 になる。

確率変数のある値に対する生起確率は、その頻度（何回起こったか）を試行数で割ったものなので、負にはなり得ません。また、確率変数の取り得るすべてを考えれば、そのどこかには必ず生起しているはずですので、後者が成り立つことも容易に理解できると思います。考える対象によって、確率分布は色々な形をとることになりますが、統計学においては特に次の一様分布、二項分布と正規分布が良く現れてきます。

2.1　一様分布

一様分布を例に、基本的な用語を確認しておきたいと思います。コインを投げて表裏を当てるゲームを考えてください。細工されていない理想的なコインでは、表と裏の出る回数は同じであると想像できます。サイコロを投げて出る目の数を当てるゲームではどうでしょう。同じく細工されていない理想的なサイコロでは、目の出る回数は同数となるでしょう[*1]。ここで「コインの表裏」「サイコロの目」は、**確率変数**（もしくは単に**変数**）と言い、試行回数（コインやサイコロを投げた数）に対するコインの表裏やサイコロの目の出る割合を**生起確率**（もしくは単に**確率**）と言います。確率変数の値を横軸にそれを変えながら生起確率をプロットしてい

* Electronic address: haga@tokushima-u.ac.jp

[*1] 少ない試行回数ではコインの表裏やサイコロの目の出る割合にばらつきが生じますが、何度も何度も繰り返していくと、コインの表裏やサイコロの目の出る回数は同じになっていきます。

くと**確率分布**を与えるグラフを書くことができます（図 1）。上記の例の確率変数は離散的ですが、連続変数でも確率分布を表現することは可能です。確率変数が連続的である場合には、確率分布を**確率密度**と呼ぶことがあります。

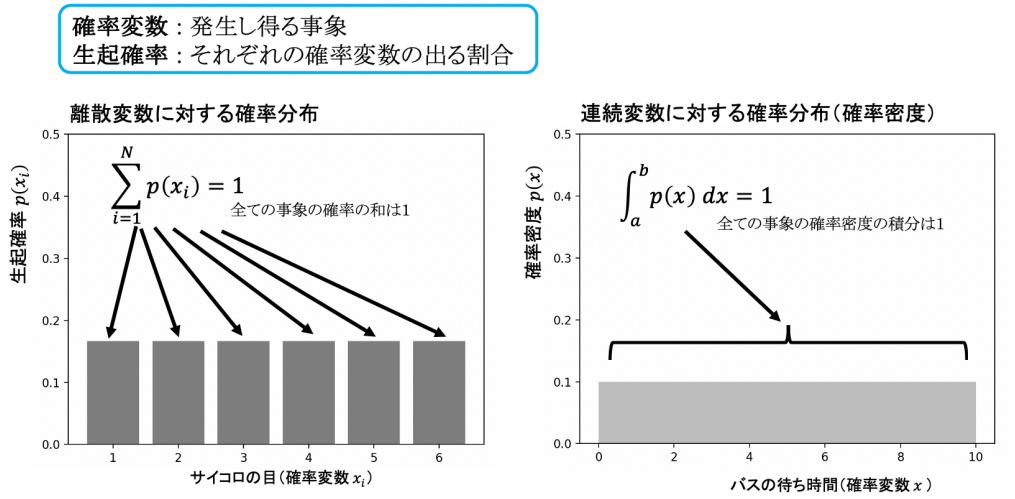

図 1　一様分布：サイコロの目の出る確率（左）と 10 分間隔で出発するバスに対する待ち時間（右）

例題：範囲 $[a, b]$ の一様分布における $a + b/2$ における確率はいくらか。

例題：サイコロを 100 回投げたときの出た目のヒストグラムを描いてみよう。
0 から 6 までの実数をランダムに 100 個取り出す → runif(100,0,6)
切り上げで整数にする → ceiling(runif(100,0,6))

```
> saikorome <- ceiling(runif(100,0,6))
> hist(saikorome,breaks=seq(0.5,6.5,1),probability=T,ylim=c(0,1))
```

2.2　二項分布

コイン投げのように、生起するパターンが 2 つ（二値）しかない事象（コイン投げでは表か裏か）について今度は複数回試行することを考えてみたいと思います。コイン投げ 1 回では表と裏の出る確率はともに 1/2 でしたが、今度は、2 回投げて 2 回とも表が出る確率、1 回だけ表が出る確率、一度も表が出ない確率（つまり裏が 2 回出る確率）はそれぞれいくつになるでしょうか？ 10 回の試行の場合はどうでしょうか？表と裏の出る確率が厳密に一致する（つまりそれぞれ 1/2 の確率で生起する）場合、N 回の試行中、x 回だけ表（もしくは裏）の出る確率は、その場合の数を数えることで得ることができます。2 回のコイン投げの場合、[表, 表]、[表, 裏]、[裏, 表]、[裏, 裏] の 4 つの組み合わせがあり、そのうち表が 2 回、0 回となる組み合わせが 1 つなので、2 回投げて 2 回とも表が出る確率、一度も表が出ない確率はともに 1/4 になります。表が 1 回出る組み合わせは 2 つあるので、2 回投げて少なくとも 1 回表となる確率は 2/4=1/2 となります。10 回コイン投げをする場合も（面倒ですが）同様に数えることで表が x 回出る確率を計算できます。$N = 2$ と $N = 10$ の場合の生起確率の分布を図 2 左上と真ん中上に示します。コイン投げの表裏の出方が等しくない場合の分布も考えることができ、表が出る確率が 0.8 の時に 10 回コインを投げた時の表の出る回数の確率を図 2 右上に示します。

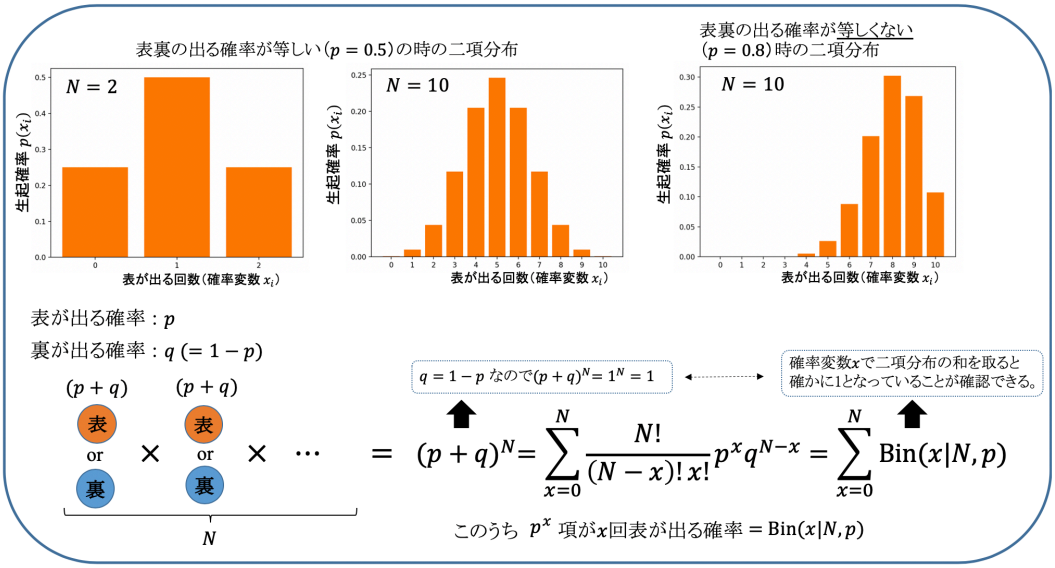

図 2　二項分布：表と裏の出る確率が等しく且つ $N = 2$ と $N = 10$ の場合の確率分布（左上および真ん中上）と表の出る確率が 0.8（裏：0.2）で $N = 10$ の場合の確率分布（左上）（右）。二項分布は二項係数で計算することができる（下）。

コイン投げのように、生起するパターンが 2 つしかない事象を試行することを**ベルヌーイ試行**と言います[*2]。そのベルヌーイ試行を N 回繰り返し、そのうち x 回表（もしくは裏）になる事象に対する x に関する確率分布（図 2 上）が**二項分布**です。今、表が出る確率が p（裏が出る確率 $q = 1 - p$）であるコインを考えます（p は 1/2 でなくて構いません）。その場合、二項分布 $\mathrm{Bin}(x|N,p)$ は

$$\mathrm{Bin}(x|N,p) = \frac{N!}{(N-x)!x!} p^x (1-p)^{N-x} \tag{1}$$

となります（図 2 下）。確率変数 x は、$x \in 0, \cdots, N$ を取ることができ、取り得る x にわたる確率の総和は確かに 1 となります。また、二項分布の平均と分散がそれぞれ Np、$Np(1-p)$ となることも、単純な計算で直接示すことができます。

例題：表と裏の出る確率が等しい（$p = 0.5$）の時の $N = 10$ の二項分布において表が 5 枚出る確率はいくらか。また、表の出る確率が 0.8 の時、$N = 10$ の二項分布において表が 10 枚出る確率はいくらか。
（R での求め方）dbinom(5,10,0.5) と dbinom(10,10,0.8) でもとまる。

例題：表と裏の出る確率が等しい（$p = 0.5$）の時の $N = 10$ の二項分布において表が **5 枚以上**となる確率はいくらか。
（R での求め方）1-pbinom(5,10,0.5) もしくは pbinom(5,10,0.5, lower.tail=F) でもとまる。

例題：上記例題の二項分布から 100 個サンプルした時のヒストグラムを描いてみよう。
二項分布からランダムに 100 個取り出す → rbinom(100,10,0.5)

[*2] コイン投げを例に取ると表と裏の出る確率は常に等しいものと考えてしまいがちですが、一般には異なる確率（つまり表と裏の生起確率がどちらかに偏っている）でも、取り得るパターンが 2 つであればベルヌーイ試行と言います。コイン投げのほか、当選・落選、合格・不合格、病気・健康、生存・死亡、買う・買わない、有り・無し、などが例としてあげられます。

```
> sample <- rbinom(100,10,0.5)
> hist(sample,breaks=seq(-0.5,10.5,1),probability=T,ylim=c(0,1))
```

2.3 正規分布（ガウス分布）

連続変数の確率分布の中で最も重要な分布が**正規分布（ガウス分布）**と呼ばれる分布です。この正規分布は社会科学・自然科学問わず、あらゆる分野・場面で頻繁に顔を出す分布であり、後半で述べる統計学においても重要な役割を果たします。身長や血圧の分布は正規分布の形をしており、対数を取ると正規分布となるような測定値も数多く存在します。前節で紹介した二項分布において N が大きいとき、正規分布に近似できることもよく知られています。

正規分布 $\mathcal{N}(x|\mu,\sigma)$ は、確率変数 x とその平均 μ と標準偏差 σ で表現され、

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

という形をもちます[*3]。これは確率変数の取り得る範囲 $x \in [-\infty, +\infty]$ で積分すると 1 となるように正規化されています。また、平均が 0、標準偏差が 1 である正規分布 $\mathcal{N}(x|0,1)$ は**標準正規分布（z 分布）**と呼ばれています。確率変数 x を $z = \frac{x-\mu}{\sigma}$ に変換することで、どのような正規分布も標準正規分布 $\mathcal{N}(z|0,1)$ に変換することができ、逆に、z を $x = \mu + \sigma z$ と変換することで、標準正規分布から平均 μ と標準偏差 σ をもつ任意の正規分布 $\mathcal{N}(x|\mu,\sigma)$ に戻すことが可能です。この性質のため、標準正規分布の性質を調べることが非常に大事です。標準正規分布の性質を理解することで、どのような平均 μ や標準偏差 σ を持っている正規分布の性質も簡単に導くことでできるからです。例えば、標準正規分布では以下の性質が成り立ちます。

- 確率変数 $z \in [-1,1]$ の区間に全体の 68.3% が入る（z による区間 $[-1,1]$ の積分値 0.683）
- 確率変数 $z \in [-2,2]$ の区間に全体の 95.4% が入る（z による区間 $[-2,2]$ の積分値 0.954）
- 確率変数 $z \in [-3,3]$ の区間に全体の 99.7% が入る（z による区間 $[-3,3]$ の積分値 0.997）
- $\cdots$

これを $x = \mu + \sigma z$ と変換して確率変数 x で同じことを記述すると、

- 確率変数 x は区間 $[\mu-\sigma, \mu+\sigma]$ に全体の 68.3% が入る（x による区間 $[\mu-\sigma, \mu+\sigma]$ の積分値 0.683）
- 確率変数 x は区間 $[\mu-2\sigma, \mu+2\sigma]$ に全体の 95.4% が入る（x による区間 $[\mu-2\sigma, \mu+2\sigma]$ の積分値 0.954）
- 確率変数 x は区間 $[\mu-3\sigma, \mu+3\sigma]$ に全体の 99.7% が入る（x による区間 $[\mu-3\sigma, \mu+3\sigma]$ の積分値 0.997）
- $\cdots$

ということになります（図 3）。これは μ を中心に $\pm\sigma$ の範囲内に確率変数 x が存在する割合（%）は 68.3%、$\pm 2\sigma$ の範囲内に確率変数 x が存在する割合（%）は 95.4% 等々、ということを意味しています。

これと同じことなのですが、今度は割合（%）を指定して確率変数の範囲を調べてみましょう（図 4）。特によく現れるのが以下に示す確率変数 95% と 99% が入る範囲です。

- 全体の 95% が入る確率変数 z の範囲は $[-1.96, 1.96]$

[*3] 2 つのパラメータ平均 μ と標準偏差 σ で正規分布の形が完全に決まり、この 2 つのパラメータを正規分布の十分統計量と言います。

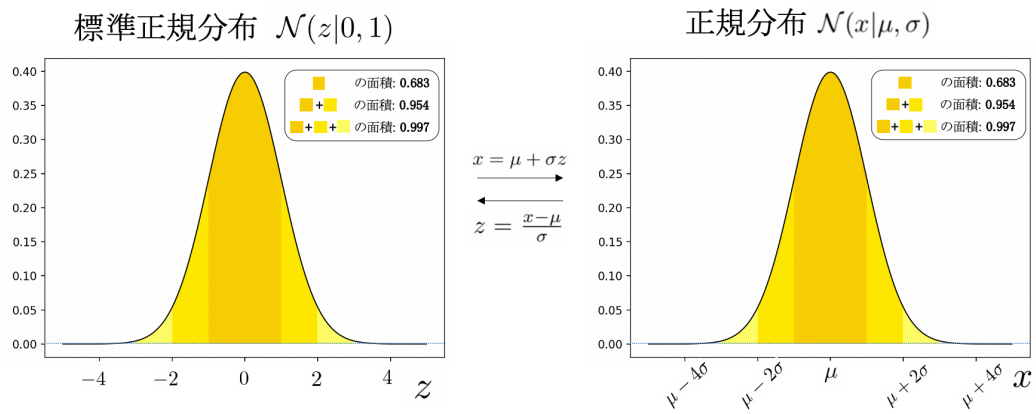

図 3　標準正規分布（左）と正規分布（右）：標準正規分布は平均 0、標準偏差 1 を持つ一方、一般の正規分布は平均 μ、標準偏差 σ を持ちます。平均値周りに釣鐘型状にばらついた分布（ばらつき具合が標準偏差で与えられます）です。

- 全体の 99% が入る確率変数 z の範囲は $[-2.58, 2.58]$

  ⇕
- 全体の 95% が入る確率変数 x の範囲は $[\mu - 1.96\sigma, \mu + 1.96\sigma]$
- 全体の 99% が入る確率変数 x の範囲は $[\mu - 2.58\sigma, \mu + 2.58\sigma]$

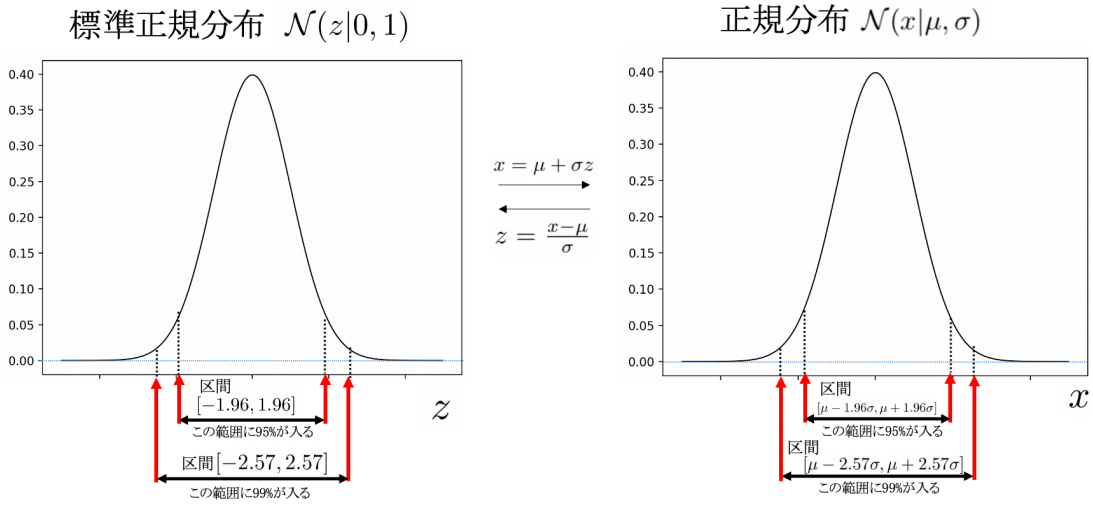

図 4　標準正規分布（左）と正規分布（右）における 95% と 99% の頻度が入る区間。

正規分布に従う確率変数をランダムに選んでみると、そのほとんどが平均値 μ 周りになるでしょう。もう少し具体的に言うと、図 4 のような正規分布に従う確率変数 x を適当に選ぶと、100 回中 95 回は $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ の範囲の変数値が選ばれる（100 回中 99 回は $[\mu - 2.58\sigma, \mu + 2.58\sigma]$ の範囲の変数値が選ばれる）ということを意味しています。この見方は後の章の統計的推論のところで再訪することになります。

正規分布が特別な分布であることを感じることができる例をここで紹介しておきましょう。物理学の熱力学という分野にエントロピー増大の法則というものがあります。エントロピーは必ず増える方向に進み、自然はエントロピーが最大になる状態を好むということを表現した法則です。エントロピーとは、簡単に言えば「乱雑さ」を測る指標であり、エントロピーが増えるということは乱雑さが増大することを意味します[*4]。情報科学では**情報エントロピー**という量によりデータの乱雑さから情報量を数値化します[*5]。面白いことに、平均 μ と標準偏差 σ を持つあらゆる分布の中でエントロピーを最大にする分布が正規分布となることを証明することができます。後の章で学ぶ中心極限定理でも正規分布が現れます。正規分布は自然界に選ばれた確率分布であり自然が最も好む神秘的な分布なのです。

3 分布の描画と値の取得

R では様々な確率分布を発生させる関数が用意されています。以下に代表的な関数を示します。

分布	関数	使用例
一様分布	dunif(x, x_{min}, x_{max})	dunif$(x, 1, 6)$, x_{min}, x_{max} はそれぞれ x の取り得る最小値, 最大値
二項分布	dbinom(x, n, p)	dbinom$(x, 10, 0.5)$, n は試行回数, p は表が出る確率
正規分布	dnorm(x, μ, σ)	dbinom$(x, 10, 1)$, μ は平均, σ は標準偏差
ポアソン分布	dpois(x, λ)	dpois$(x, 1)$, λ はポアソン分布のパラメータ λ のこと
カイ二乗分布	dchisq(x, df, ncp)	dchisq$(x, 5)$, df はカイ二乗分布の自由度、ncp は非中心化パラメータ
t 分布	dt(x, df, ncp)	dt$(x, 5)$, df は t 分布の自由度、ncp は非中心化パラメータ
F 分布	df$(x, df1, df2, ncp)$	df$(x, 4, 8)$, $df1, df2$ はそれぞれ分子・分母のカイ二乗分布の自由度
ガンマ分布	dgamma(x, a, b)	dgamma$(x, 2, 0.5)$, a, b はガンマ分布のパラメータ
ロジスティック分布	dlogis(x, μ, s)	dlogis$(x, 5, 1)$, μ, s はロジスティック分布のパラメータ

```
> x<- seq(0,10,by=0.1)
> plot(x, dnorm(x,5,1), xlim=c(0,10), ylim=c(0,0.5), type="l", col="blue")
> par(new=T)
> plot(x, dnorm(x,3,1), xlim=c(0,10), ylim=c(0,0.5), type="l", col="red")
> par(new=T)
> plot(x, dnorm(x,5,3), xlim=c(0,10), ylim=c(0,0.5), type="l", col="green")
```

また、上と同じグラフ化は curve 関数を使って curve(dnorm$(x, 5, 1)$,0,10) とするだけでも行えます。

dnorm(x, μ, σ) は x を指定した時の正規分布の値を返してくれる関数です。この他にも R には便利な関数が用意されています。例えば正規分布の**信頼区間**を与える変数 x の値は qnorm(p, μ, σ) で求めます（図 5)。ここで p に 0.025 を与えると 95% 信頼区間の下限、0.975 を与えると 95% 信頼区間の上限がそれぞれ得られます（両側の裾野を合わせて 5% となるので 0.025 と $0.975(= 1 - 0.025)$ を入れます)。今度は逆に上限 q を与えた時にその範囲 $[-\infty, q]$ にある分布の確率（分布と x 軸で囲まれる面積）を求めてみます。これは pnorm(q, μ, σ) で与えられます。例えば標準正規分布（平均 0、標準偏差 1）において $q = -1.96$ とした時に

[*4] 自分の部屋が汚くなるのは自然が好むから？

[*5] 離散変数では $-\sum_i p(x_i) \ln p(x_i)$、連続変数では $-\int p(x) \ln p(x) dx$ がエントロピー（連続変数では微分エントロピーという）であり、情報エントロピーはビット演算の都合上底を 2 にしたものです。滅多に起こり得ない稀な現象=情報が一杯詰まっている → 情報エントロピーは小さい、という関係があります。画像がノイズだらけ（エントロピーが大）だと何が写っているのかわからない（情報が小）ということです。

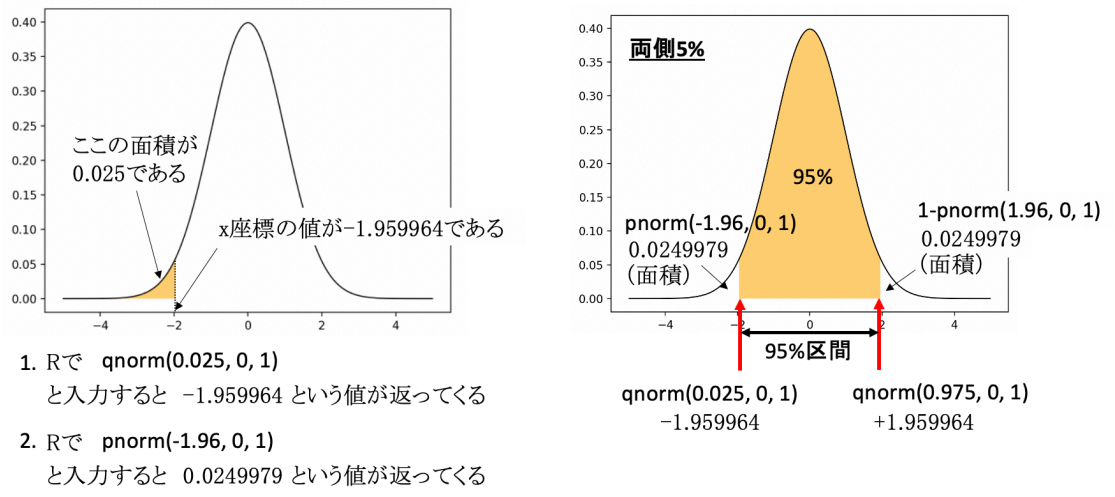

図 5　標準正規分布（左）と正規分布（右）における 95% と 99% の頻度が入る区間。

pnorm$(q, 0, 1)$ で返ってくる値はいくつになるでしょうか？これは 0.0249979 となります（ほぼ 0.025）。つまり 95% 信頼区間の下限が約 -1.96 であるということを裏返して求めたことになります。任意の値 q に対して pnorm(q, μ, σ) で得られる値は後述する p 値と関係しています。例えば標準正規分布において $-1 < x < +1$ の範囲外となる確率は pnorm$(-1, \mu, \sigma) + (1-\text{pnorm}(1, \mu, \sigma)) \sim 0.317$ で得ることができます。これが後に学ぶ p **値**に相当します。

他の分布でも同じような使い方で分布の値や信頼区間、p 値などが計算できます。以下に代表的な確率分布の R における仕様をまとめておきます。“d”の部分を “q”に変更すると p 値に対応する確率変数の値、“p”に変更すると確率変数の値に対応する p 値がそれぞれの分布で得られることになります。遊んでみてください。

4　演習

1. 正解の確率が 0.8 である問題を 10 人に解かせたときに全員が正解する確率を求めなさい。
2. 上の問題で 5 人以下である確率を求めなさい。
3. 100 人中 1 人が感染症にかかっている状況で、10000 人を集めた場合、10 人以上が感染症にかかっている確率を求めなさい。
4. 上の問題で 10 人未満の確率を求めなさい。
5. 平均点が 50 点で標準偏差が 10 点で正規分布となるテストがある。この正規分布を R で図示するコマンドをかけ。
6. 上の正規分布で上側 2.5% となるためには点数以上取らなければならないか。
7. 引き続き上の正規分布で、75 点以上取った人の割合（確率）を求めなさい。
8. 自由度 $n = 10$ の t 分布を R で図示しなさい。
9. 上の t 分布で上側 2.5% となる t 値を R で求めなさい。
10. 自由度 1 の χ 二乗分布で $\chi^2 \geq 1$ となる確率を求めなさい。

manaba のレポートにおいて、R 上で実行した内容と出力結果とともに解答（説明）すること。