

Rと推定

1. 母集団と標本

国勢調査のような調査対象に対して漏れなく行う調査を**全数調査**と言い、アンケート調査、テレビの視聴率調査のような調査対象の集団から一部を抽出して行う調査を**標本調査**と言う。このとき、調査対象全体を**母集団**、全体から一部を取り出した部分を**標本**、あるいは**サンプル**と呼び、取り出した個数を**標本の大きさ**、あるいは**標本サイズ**と呼ぶ。

標本調査の場合は、標本のデータが母集団の性質をなるべく忠実に反映するように標本を抽出しなければならない。そのため、標本抽出には、母集団を構成する要素が偏りなく均一の確率で抽出されるような抽出方法を用いる。このような抽出方法を**無作為抽出法**と呼ぶ。

標本調査では、しばしば標本データの**統計量**（比率、平均、分散など）を用いて母集団の特性値（比率、平均、分散など）を推測する。母集団の特性値比率、平均、分散などを**母数**（母比率、母平均、母分散など）と呼ぶ。標本データの統計量を用いた母数の推定は確率分布に基づいて行う。

2. 確率変数と確率

確率変数を X とし、1つの値 $X = b$ が与えられた場合、確率 $P(X \leq b)$ を**下側確率**と呼ぶ。連続型確率変数を例として図で表すと図1のような確率密度曲線と横軸との間に囲まれた b までの面積が確率 $P(X \leq b)$ である。

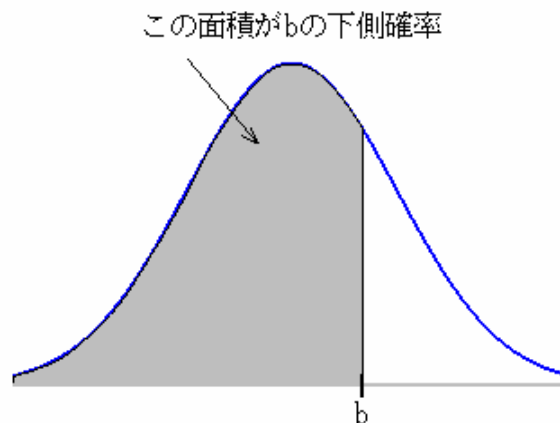


図1 連続型確率変数の下側確率 $P(X \leq b)$

R には下側確率の値を求める関数が用意されている。正規分布の下側確率を求める関数は `pnorm` である。例えば、標準正規分布の 2 の下側確率は次のように求める。

```
> pnorm(2,mean=0,sd=1)
[1] 0.9772499
```

標準正規分布の場合は、引数 `mean=0`、`sd=1` を省略してもよい。mean は平均、sd は標準偏差である。下側確率 $P(X \leq b)$ に対応する $P(X \geq b)$ を上側確率と呼ぶ。上側確率は全体から下側確率を引くことで求めることができる。

$$P(x \geq b) = 1 - P(x \leq b)$$

図2 に標準正規分布の下側確率 $P(X \leq 2)$ と上側確率 $P(X \geq 2)$ を示す。

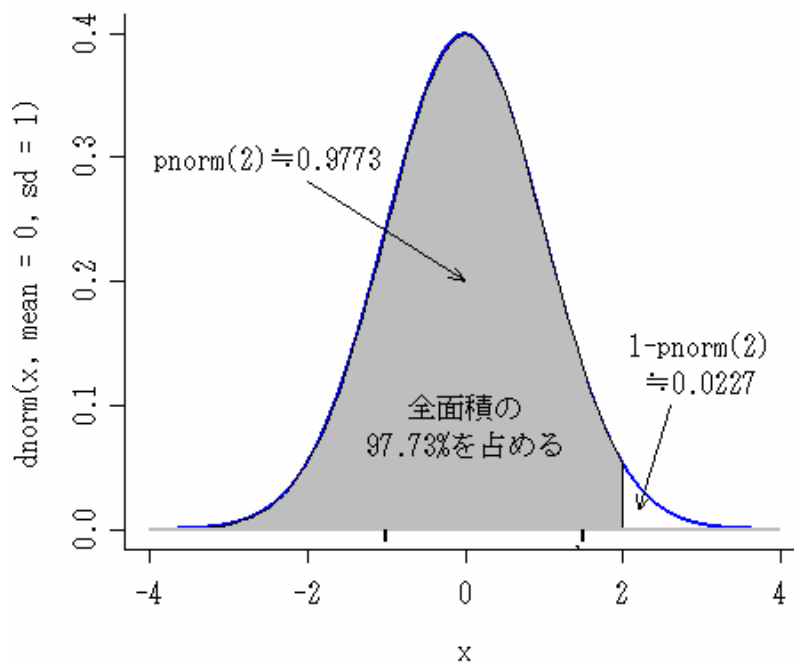


図2 標準正規分布の $P(X \leq 2)$

確率分布の曲線が原点を中心とした対称分布である場合、確率 $P(-a \leq X \leq a)$ を**両側確率**と言う。区間 $[a, b]$ の確率 $P(a \leq X \leq b)$ は図 3 のような区間 $a \leq X \leq b$ 内の密度曲線と横軸との間に囲まれた面積で、次のように求める。

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

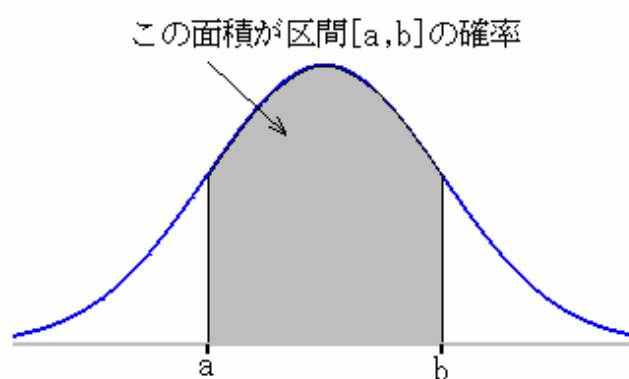


図 3 確率 $P(a \leq x \leq b)$

例えば、標準正規分布の $[-2, 2]$ の確率は次のように求める。

```
> pnorm(2) - pnorm(-2)
[1] 0.9544997
```

標準正規分布の $[-2, 2]$ の確率は約 0.96 である。これは -2 から 2 までの面積は全体の約 96% を占めることを意味する。

図 3 のような分布では、確率密度曲線の峰(中心)の近隣部分は試行を繰り返し行う際に、観測の結果が現れる確率が高く、峰から両側に離れるほど確率が低い。データを分析する際には、確率が非常に低いものはしばしば無視する。問題は、確率が低いか、それとも高いかは何を基準とするかである。その基準として確率 $P(a \leq X \leq b)$ が用いられている。一般的には、確率

$P(a \leq X \leq b)$ を 0.9、0.95、0.99 のように決めておき、それに対応する a, b を求め、 a より小さく、かつ b より大きい確率変数の確率は低いと判断する。この a 、 b を**分位点(quantile)**と呼ぶ。正規分布の分位点は正規分布の分位点関数 `qnorm` を用いて求めることができる。分位点関数は下

側確率関数の逆関数である。

例えば、標準正規分布の 0.975 の分位点は次のように求める。

```
> qnorm(0.975)
[1] 1.959964
```

図 4 に標準正規分布の分位点、下側確率の対応関係を示す。

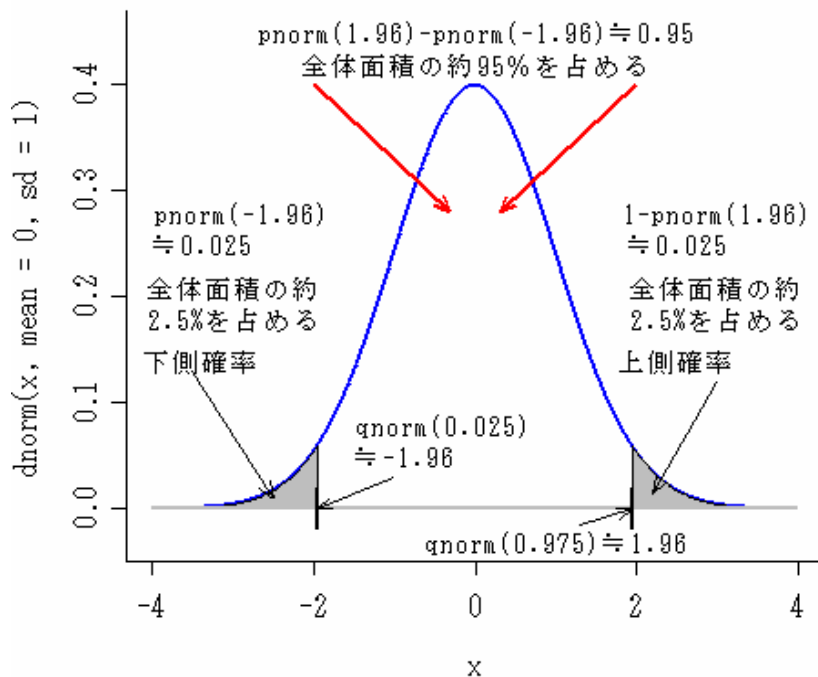


図 4 標準正規分布の分位点と下側確率

標準正規分布は $X = 0$ を中心とした左右対称分布である。 $X = 0$ を中心とした確率 $P(a \leq X \leq b)$ が 0.9、0.95、0.99 の分位点 a 、 b の対応関係を表 1 に示す。

表 1 標準正規分布の確率と分位点

$P(a \leq X \leq b)$	a	b
0.9	-1.64485	1.64485
0.95	-1.95996	1.95996
0.99	-2.57583	2.57583

3. 推定

アンケート調査や製品の検査などでは全数調査が不可能な場合がほとんどである。このような

標本調査では標本データの統計量を用いて、その標本が属する母集団の母数を推測する。標本データの統計量を用いて母数を推測することを**推定**と言う。推定には点推定と区間推定がある。

点推定とは、標本の統計量を母数と見なす推定方法である。これは、標本サイズが十分大きい場合は標本分布がその母集団の分布に近似するという考えに基づいている。しかし、標本サイズが十分大きくない場合は、同一の母集団から抽出した異なる標本の統計量はそれぞれ異なる。区間推定は、この異なる値が納められる範囲・区間を推定することである。区間推定には、いろいろな統計量について推定することが可能であるが本稿では母平均と母比率の区間推定のみについてシミュレーションを通じて説明する。

3.1 データの標準化

まず一つのシミュレーションを行うことにする。平均が 170 で標準偏差が 5 である正規分布 $N(170,5^2)$ の乱数を 300 個発生させる。これは平均が 170cm、標準偏差が 5 である母集団から 300 人を無作為に抽出したと考えることもできる。

```
>X<-rnorm(300,170,5)
```

発生させた乱数 X について、次に示した式の変換を行い、 Z の平均と分散を求めて見よう。

$$Z = \frac{X - \bar{X}}{\sqrt{V}} = \frac{X - \bar{X}}{s}$$

```
> Z<-(X-mean(X))/sqrt(var(X))
```

```
> mean(Z)
```

```
[1] -1.436773e-15
```

```
> var(Z)
```

```
[1] 1
```

上記の-1.436773e-15 は-1.436773 の小数点を左に 15 桁を移動した値に等しいので 0 であると見なしてもよい。この値は乱数データに基づいた計算結果であるので、読者が同じのコマンドを実行してもこれと同じの結果が得られないが、変換されたデータの平均は 0、標準偏差は 1 に近似する点では一致する。

3.2 標本平均の性質

前節のような同じの正規分布の乱数を繰り返し発生させ、毎回発生させた乱数の平均値の平均と分散を考察してみよう。

平均が 170、標準偏差が 5 である正規分布 $N(170,5^2)$ から 300 の乱数を発生させたデータを 1 つの標本とし、その平均値を求める。このような乱数を 1000 回発生させると 1000 個の平均値が

得られる。この 1000 個の平均値の平均と分散（あるいは標準偏差）の規則性について注意して欲しい。

次に標本サイズ 300 の乱数を 1000 回発生させた標本平均の平均と分散を求めるコマンドと結果を示す。

```
> kekka<-matrix(0,1000,300)
>for(i in 1:1000){kekka[i,]<-rmorm(300,170,5)}
> temp<-apply(kekka,1,mean)
> mean(temp)
[1] 169.9892
> var(temp)
[1] 0.08373682
```

求めた標本平均の平均 169.9892 は母集団の平均 170 に近似し、標本平均の分散 0.08373682 は母集団の分散 25 を標本サイズ 300 で割った値 $25/300=0.083333335$ に近似している。これは次の定理のシミュレーションである。

定理: 正規分布 $N(\mu, \sigma^2)$ から抽出した標本サイズが n で
 $N(\mu, \frac{\sigma^2}{n})$ に従う。

上記のシミュレーションでは $N(170, 5^2)$ の乱数を発生している。よって発生された乱数の平均は $N(170, \frac{5^2}{300})$ に従う。この結果と前節の標準化に関する結果を用いると、標本平均を標準化した Z は標準正規分布に従うことが導かれる。

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \text{とき} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

表記の中「 \sim 」は「従う」ことを意味する。ここでは母集団の平均と標準偏差を用いている。しかし、実際の問題では母集団の標準偏差が未知の場合が多い。標本のサイズが大きい場合は標本の不偏分散 V (標準偏差 $s = \sqrt{V}$) を母分散 σ^2 の替りに用いることもできる。

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \text{とき} \quad Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \approx N(0,1)$$

式の中の記号「 \approx 」は近似的に従うことを意味する。問題は、標本のサイズがどのぐらいであ

れば大きいといえるかである。経験則としては 30 以上であれば大標本といわれているので一つの目安となる。

標本サイズが小さいときには標準化された確率変数は自由度 $n-1$ の t 分布に従うことが知られている。

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{とき} \quad T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

3.3 区間推定

確率変数 Y の確率 $P(a \leq Y \leq b)$ が 90%、95%、99% のように与えられたときそれに対応する区間 $[a, b]$ を信頼係数 0.9=90%、0.95=95%、0.99=99% の信頼区間と呼ぶ。信頼区間を求めることを区間推定という。信頼係数 0.95=95% で求めた信頼区間をイメージ的に説明すると、100 回の試行を行った時、95 回の結果は信頼区間内に納めるが、5 回ぐらいの結果は信頼区間 $[a, b]$ 内に納めることが期待できない。信頼係数は $1-\alpha$ あるいは $100(1-\alpha)\%$ で表し、 α を有意水準と呼ぶ。図 5 に標準正規分布における有意水準、信頼係数、信頼区間などの対応関係を示す。

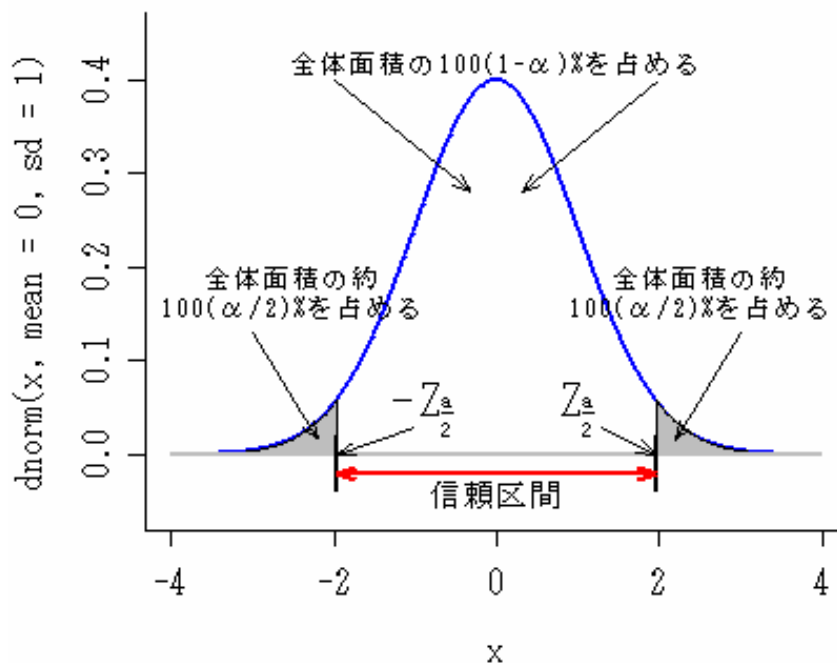


図 5 標準正規分布の信頼区間

3.3.1 母平均の信頼区間

前節の $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ に基づいて導出した母平均の信頼区間の関係式を次に示す。

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\ &= P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq z_{\alpha/2}) \\ &= P(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \end{aligned}$$

この式の中の次に示す不等式が与えている区間が有意水準 α (あるいは信頼係数 $1 - \alpha$) における母平均の信頼区間である。

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

不等式の中の $z_{\alpha/2}$ は有意水準 α が具体的に与えると正規分布の分位点関数 `qnorm` を用いて求めることができる。 \bar{X} 、 n 、 σ はそれぞれ標本の平均、サイズ、母集団の標準偏差である。よって、母集団の標準偏差が既知である場合は、母平均の信頼区間を簡単に求めることができる。母集団の標準偏差が未知であっても標本サイズが大きい場合は、標本の不偏分散を母分散の代わりに用いることが可能である。

標本サイズが小さい場合は、母分散 σ^2 の代わりに標本の不偏分散を用いると次の確率変数は自由度 $n-1$ の t 分布に従うことが知られている。

$$T = \frac{\bar{X} - \mu}{\sqrt{V/n}} \sim t(n-1)$$

よって、標本サイズが小さい場合、標本の平均と分散を用いて母平均の信頼区間を求めるときには次の式を用いる。

$$\bar{X} - t\left(\frac{\alpha}{2}, n-1\right) \sqrt{\frac{V}{n}} \leq \mu \leq \bar{X} + t\left(\frac{\alpha}{2}, n-1\right) \sqrt{\frac{V}{n}}$$

$t\left(\frac{\alpha}{2}, n-1\right)$ は **R** では t 分布の分位点関数 `qt` を用いて求めることができる。例えば、有意水準 $\alpha = 0.05$ 、標本サイズ $n = 10$ の $t\left(\frac{0.05}{2}, 10-1\right)$ は次のように求める。

```
> qt(0.025,9)
```

```
[1] -2.262157
```

このように R で求めた $t(\frac{\alpha}{2}, n-1)$ には正負の符号がついているので、区間の端点を計算する際には、式

$$\text{左の端点: } \bar{X} - t\left(\frac{\alpha}{2}, n-1\right) \sqrt{\frac{V}{n}}$$

$$\text{右の端点: } \bar{X} + t\left(\frac{\alpha}{2}, n-1\right) \sqrt{\frac{V}{n}}$$

のなかの $t(\frac{\alpha}{2}, n-1)$ は絶対値を用いるべきである。

3.3.2 母比率の信頼区間

試行、実験、調査などで、ある観測項目が現れるか、現れないかに関する結果は二項分布 $B(n, p)$ に従う。確率変数 k が 1、2、3、・・・、30 をとる二項分布 $B(50, 0.3)$ のグラフを作成し、さらに同じ座標上で平均が $15=50*0.3$ 、分散が $10.5=50*0.3*0.7$ である正規分布のグラフを作成するコマンドを次に示しその結果を図6に示す。

```
> x<-0:30  
> plot(x,dbinom(x,50,prob=0.3),type="h")  
> sd1<- sqrt(50*0.3*0.7)  
> curve(dnorm(x,mean=0.3*50, sd=sd1,add=T)
```

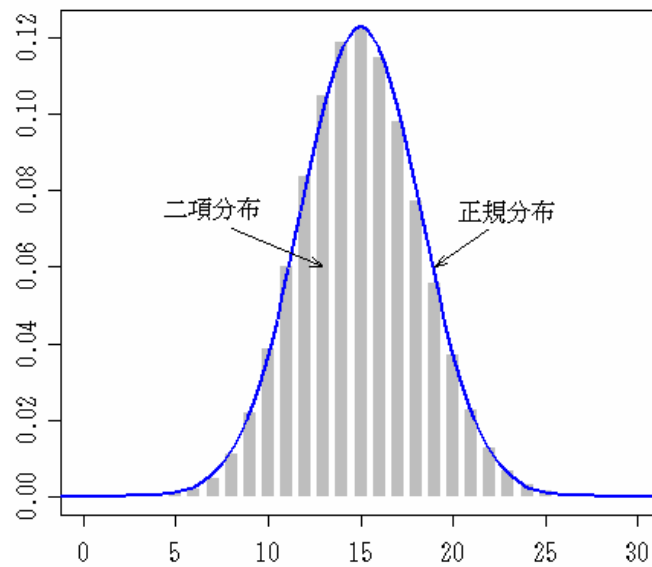


図6 二項分布と正規分布

図6から標本サイズが大きい場合(ここでは50)二項分布は正規分布に非常によく近似していることがわかる。つまり標本サイズが大きいときには二項分布 $B(n, p)$ は平均 np 、標準偏差 $\sqrt{np(1-p)}$ の正規分布に近似的に従い、さらにそれを標準化すると標準正規分布に近似的に従う。

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1)$$

この性質を用いると比率の推定区間は次の式を用いて求めることができる。

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\ &= P(-z_{\alpha/2} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq z_{\alpha/2}) \end{aligned}$$

式の中の不等式を次のように整理することができる。式の中の P_a は母比率で、 \hat{p} は標本の比

率 $\frac{x}{n}$ である。

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p_a \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

最近テレビや新聞では頻繁に政権の支持率等に関する調査データが用いられるようになった。そのような情報の受け取る際には正しい統計知識が必要である。

例えば、1000 人に対して調査を行った結果、現政権に対する支持率が 45%だとする。有意水準 5% ($\alpha=0.05$) の場合の母比率の信頼区間の計算結果を次に示す。

```
> z<-abs(qnorm(0.025))
> 0.45-z*sqrt(0.45*0.55/1000)
[1] 0.4191656
> 0.45+z*sqrt(0.45*0.55/1000)
[1] 0.4808344
```

得られた結果を小数点 4 桁まで丸めると母比率の推定区間は $0.4192 \leq p_a \leq 0.4808$ となる。この推定区間は、今回の調査では支持率が 45% という結果が得られているが、母集団の支持率はおおよそ 42%~48% であると推測されることを意味する。

表 1 に頻繁に使用されている確率分布の下側確率と分位点を求める R の関数を示す。

表 1 R における下側確率と分位点の関数

分布の名	下側確率	分位点
一様(Uniform)分布	<code>punif(q, min=0, max=1, ...)</code>	<code>qunif(p, min=0, max=1, ...)</code>
二項(Binomial)分布	<code>pbinom(q, size, prob, ...)</code>	<code>qbinom(p, size, prob, ...)</code>
ポアソン(Poisson)分布	<code>ppois(q, lambda, ...)</code>	<code>qpois(p, lambda, ...)</code>
正規(Normal)分布	<code>pnorm(q, mean=0, sd=1, ...)</code>	<code>qnorm(p, mean=0, sd=1, ...)</code>
カイ2乗(Chi-square)分布	<code>pchisq(q, df, ncp=0, ...)</code>	<code>qchisq(p, df, ncp=0, ...)</code>
t 分布	<code>pt(q, df, ...)</code>	<code>qt(p, df, ...)</code>
F 分布	<code>pf(q, df1, df2, ...)</code>	<code>qf(p, df1, df2, ...)</code>
ガンマ(Gamma)分布	<code>pgamma(q, shape, ...)</code>	<code>qgamma(p, shape, ...)</code>
ベータ(Beta)分布	<code>pbeta(q, shape1, shape2, ...)</code>	<code>qbeta(p, shape1, shape2, ...)</code>
対数正規(Lognormal)分布	<code>plnorm(q, meanlog = 0, sdlog = 1, ...)</code>	<code>qlnorm(p, meanlog = 0, sdlog = 1, ...)</code>
ロジスティック(Logistic)分布	<code>plogis(q, ...)</code>	<code>qlogis(p, ...)</code>
指数(Exponential)分布	<code>pexp(q, rate = 1, ...)</code>	<code>qexp(p, rate = 1, ...)</code>
負二項(Negbinomial)分布	<code>pnbinom(q, size, prob, mu, ...)</code>	<code>qnbinom(p, size, prob, mu, ...)</code>
幾何(Geometric)分布	<code>pgeom(q, prob, ...)</code>	<code>qgeom(p, prob, ...)</code>
超幾何(Hypergeometric)分布	<code>phyper(q, m, n, k, ...)</code>	<code>qhyper(p, m, n, k, ...)</code>
コーシー(Cauchy)分布	<code>pcauchy(q, location=0, scale= 1, ...)</code>	<code>qcauchy(p, location=0, scale = 1, ...)</code>
ワイブル(Weibull)分布	<code>pweibull(q, shape, scale=1, ...)</code>	<code>qweibull(p, shape, scale = 1, ...)</code>