Ｒで始めるデータサイエンス③

Ｒを使って
区間推定

1． ネジが沢山！長さは？　～区間推定
2． 長さは規格に準拠？　～標本平均
3． 区間推定の数理　～確率分布
4． Ｒで区間推定　～場合分け

ネジの長さを調べたい

物凄く沢山のネジ
（L 個）

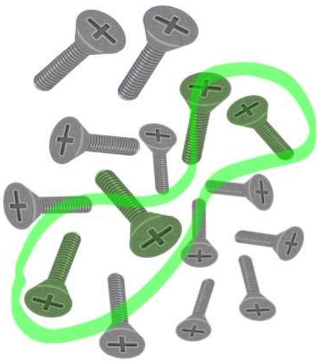
それぞれの長さは x_i
（$i = 1,2,\cdots,L$）

全部の平均＝$\frac{x_1+x_2+\cdots+x_L}{L}$
を知りたい

全部（L 個）を調査したい
けど、**多過ぎて無理！**

ネジを無作為に抽出する

物凄く沢山
全部は調査できない！



無作為に
抽出する

数個を抽出（n 個）
少数なら調査できる！

標本の平均$=\frac{19+18+20+17}{4}=18.5$

抽出するたびに 値がバラつく

物凄く沢山
全数調査は無理！

無作為
に抽出

数個を抽出（n 個）
調査できるが
値がバラバラ※

標本平均1＝19.5

標本平均2＝17.5

標本平均3＝18.5

⋮　　⋮

※値が確率的に変動する

全数調査できたら 平均 はどの 範囲？

物凄く沢山
全数調査は無理！

無作為に抽出

数個を抽出（n 個）
１回だけ標本調査する

標本の平均
標本の分散

下限＜全数の平均＜上限

95%の信頼区間
（下限と上限を知りたい）

「区間推定」にできること

物凄く沢山
全数調査は無理！

無作為に抽出

数個を抽出（n 個）
１回だけ標本調査する

標本の平均
標本の分散

区間推定にできること

下限＜全数の平均＜上限

95%の信頼区間
（下限と上限を知りたい）

「区間推定」の例

全数調査して
平均を知りたい
（分散 σ^2 は既知※）

無作為
に抽出

標本調査の結果

標本の個数 n
標本の平均 $\bar{X}$

下限は $\bar{X} - 2.0\,\sigma/\sqrt{n}$

上限は $\bar{X} + 2.0\,\sigma/\sqrt{n}$

※何故、こうなるかを学習します

区間推定の例

区間推定の結果

下限＜全数の平均＜上限

95%の信頼区間

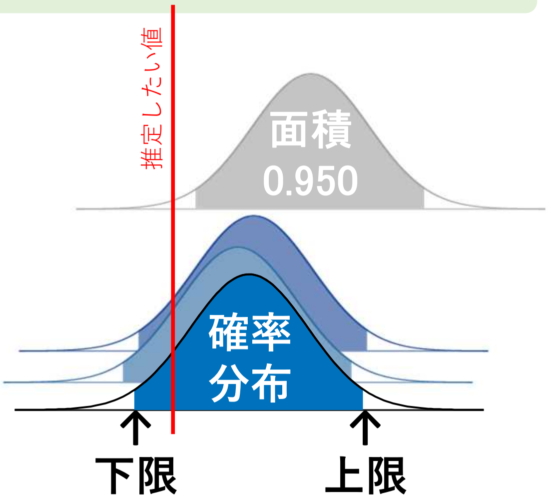
※分散が未知の場合は後ほど説明

「95%の信頼区間」とは？

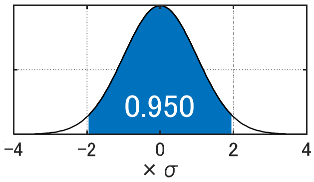
下限 ＜ 推定したい値 ＜ 上限

区間推定を100回
繰り返したら…

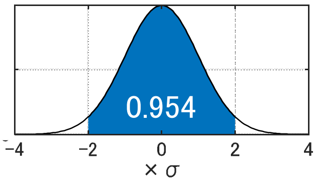
そのうち 95 回は
推定したい値が
この区間内に入る



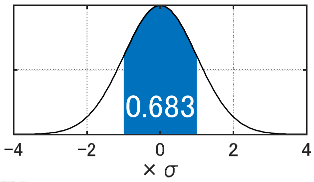
正規分布 $N(0\,,\sigma^2)$ の場合



$\pm\,1.96\,\sigma$ 　95.0 %の信頼区間



$\pm\,2.00\,\sigma$ 　95.4 %の信頼区間



$\pm\,1.00\,\sigma$ 　68.3 %の信頼区間

Ｒで始めるデータサイエンス③

Ｒを使って 区間推定

1．ネジが沢山！長さは？　～区間推定
2．長さは規格に準拠？　～標本平均
3．区間推定の数理　～確率分布
4．Ｒで区間推定　～場合分け

規格に準拠しているか？（1/4）

- 規格の下で、長さ 20 のネジを生産している
- これまでは、平均 20、標準偏差は 1 だった

※ネジの長さは正規分布すると仮定

- 機械を刷新したら、長さが変わったようだ
- 無作為に 36 個抽出したら、
  その平均が 20.5 だった（規格より 0.5 長い）
- 長さ 20 という 規格に準拠しているか？

標準偏差はこれまでと同じ 1 である

規格に準拠しているか？（2/4）

標本の個数 $n = 36$ **全数の標準偏差** $\sigma = 1$

標本の平均 $\bar{X} = 20.5$ ※ネジの長さは正規分布すると仮定

- 平均が 20.5 なので、規格より 0.5 長い
- どの 36 個を抽出するかで、平均はバラつく
- バラツキの範囲を求めよう

下限は $20.5 - 2/\sqrt{36} = 20.2$

上限は $20.5 + 2/\sqrt{36} = 20.8$

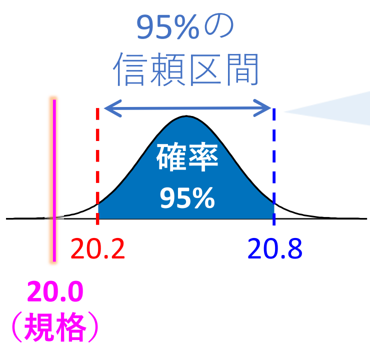
$\bar{X} - 2.0\,\sigma/\sqrt{n}$

$\bar{X} + 2.0\,\sigma/\sqrt{n}$

先ずは使ってみましょう

規格に準拠しているか？（3/4）

20.2 ＜ 全数の平均 ＜ 20.8
下限　　　　　　　　上限



規格より長くなっている
※ 抽出の数を増やせば
範囲はもっと狭くなる

規格に準拠していない

規格に準拠しているか？（4/4）

全数の長さの
平均を知りたい
（標準偏差は $\sigma = 1$）
長さは正規分布すると仮定

無作為
に抽出

標本の個数 36
標本の平均 20.5

点 として推定

20.2 ＜ 全数の平均 ＜ 20.8
95%の信頼区間

区間 として推定

下限は $20.5 - 2/\sqrt{36} =$ **20.2**
上限は $20.5 + 2/\sqrt{36} =$ **20.8**
※何故こうなるかは、後ほど説明

区間を絞って精度を上げるには？

ネジを大量生産する機械を刷新したら、
ネジが長くなった気がする。

５個を無作為に抽出して区間推定したら、
20 ± 1（mm）だった（区間幅が ± 1 ）。

区間幅を ± 0.1 未満とするには、
何個を抽出する必要があるか？

但し、標準偏差は
$\sigma = 1.12$ とする

全数の平均 $\mu = 20 \pm 1$

$\downarrow$

± 0.1 未満にしたい

区間を絞って精度を上げるには？

n 個を無作為に抽出して区間推定して、
20 ± 0.1 とするには、

$$20 + 2\sigma/\sqrt{n} < 20 + 0.1$$

但し、標準偏差は
$\sigma = 1.12$ とする

が必要なので、

$$2\sigma/\sqrt{n} < 0.1$$

$$n > \left(\frac{2 \times 1.12}{0.1}\right)^2 = 501.8$$

502 個
以上を調べる
必要がある！

効果10倍、努力100倍

下限＜全部の平均＜上限

95%の信頼区間

区間推定の結果

標本数を n とする

区間の幅は
ルート n に
反比例する

下限は $\bar{X} - 2\,\sigma/\sqrt{n}$

上限は $\bar{X} + 2\,\sigma/\sqrt{n}$

幅を 1/10 にするには
100倍の標本調査が必要

区間推定の例

Rで始めるデータサイエンス③

Rを使って 区間推定

1． ネジが沢山！長さは？ ～区間推定
2． 長さは規格に準拠？ ～標本平均
3． 区間推定の数理 ～確率分布
4． Rで区間推定 ～場合分け

信頼区間、母平均、母分散

母集団の全てを調べて
母平均 μ を知りたい
母集団は正規分布すると仮定
母分散 σ^2 は既知※

↓ 無作為に抽出

標本調査の結果
標本の個数 n
標本の平均 $\bar{X}$

→

下限は $\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}}$
上限は $\bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}$

↑

下限＜ 母平均 μ ＜上限
95%の信頼区間

※母分散が未知の場合は後ほど説明

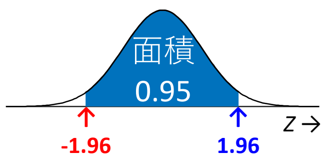
信頼区間 は 確率分布 で決まる

$$\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < 1.96\,\frac{\sigma}{\sqrt{n}} + \bar{X}$$

$$-1.96\,\frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < 1.96\,\frac{\sigma}{\sqrt{n}}$$

$$-1.96 < \boxed{\frac{\mu - \bar{X}}{\sigma/\sqrt{n}}} < 1.96$$

↑
この値の確率分布



下限＜ 母平均 μ ＜上限
95%の信頼区間

↑

下限は $\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}}$
上限は $\bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}$
母分散 σ^2 は既知のとき
X は正規分布すると仮定

データ の 標準化 とは?

平均 μ
分散 σ^2/n の
正規分布に従う

正規分布 $N(\mu, \sigma^2/n)$

確率変数 $\bar{X}$

X が正規分布するならば
$\bar{X}$ も正規分布する

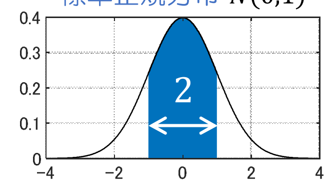
データの標準化

平均 0
分散 1 の
正規分布に従う

標準正規分布 $N(0,1)$

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$$

標準正規分布 $N(0,1)$



区間幅 が $\sqrt{n}$ に反比例する理由

n個の平均

$$E[X] = \frac{X_1+X_2+\cdots+X_n}{n} = \bar{X}$$

n個の平均の分散

$$V[\bar{X}] = V\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]$$

$$= \frac{V[X_1+X_2+\cdots+X_n]}{n^2}$$

$$= \frac{V[X_1]+V[X_2]+\cdots+V[X_n]}{n^2}$$

$$= \frac{\sigma^2}{n}$$

$$\frac{X_1+X_2+\cdots+X_n}{n} \rightarrow \mu$$

根拠
!

標本平均 $\bar{X}$ は
平均が μ
標準偏差が $\sigma/\sqrt{n}$
の**正規分布**に従う

↑
中心極限定理による
(nが大きいとき)

$\bar{X}$ の標準偏差は $\sigma/\sqrt{n}$

$\bar{X}$ は μ に収束 ($n \rightarrow$ 大)

Rで始めるデータサイエンス③

Rを使って 区間推定

1． ネジが沢山！長さは？　～区間推定
2． 長さは規格に準拠？　～標本平均
3． 区間推定の数理　～確率分布
4． Rで区間推定　～場合分け

場合分け

	母集団 全ての個体、正規分布		標本 無作為に抽出したｎ個	
	母平均	母分散	標本平均	標本分散
場合1	推定する	既知	データから計算	不要
場合2		未知	データから計算する	

	母集団×2つ※ 全ての個体、正規分布		標本×2つ 無作為に抽出したｍ個とｎ個	
	母平均の差	母分散	標本平均	標本分散
場合3	推定する	既知	データから計算	不要
場合4		未知	データから計算する	

※２標本問題と呼ばれる

場合１

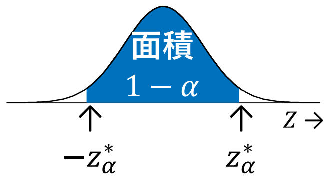
	母集団 全ての個体、正規分布		標本 無作為に抽出した n 個	
	母平均 μ	母分散 σ	標本平均 $\bar{X}$	標本分散
場合1	推定する	既知	データから計算	不要

$$-z_\alpha^* \le \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} \le z_\alpha^*$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

↑

標準正規分布 $N(0,1)$ に従う



z_α^* は両側 α 点

信頼度 $100(1-\alpha)$ %

信頼度90% → $\alpha = 0.10$ → $z_{0.10}^*$=1.64

信頼度95% → $\alpha = 0.05$ → $z_{0.05}^*$=1.96

信頼度99% → $\alpha = 0.01$ → $z_{0.01}^*$=2.58

数式表現のまとめ（1/2）

$$-z_\alpha^* \le \frac{\mu - \bar{X}}{\sqrt{\sigma^2/n}} \le z_\alpha^* \quad \Rightarrow \quad \left|\frac{\mu - \bar{X}}{\sigma/\sqrt{n}}\right| \le z_\alpha^*$$

↑ 両側 α 点　　↑ 標準化

$$\bar{X} - z_\alpha^* \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_\alpha^* \frac{\sigma}{\sqrt{n}}$$

$100(1-\alpha)$ % の信頼区間

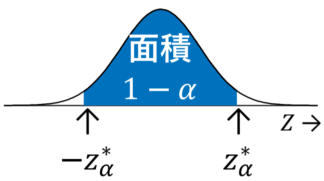
$$\mu \in \left[\ \bar{X} - z_\alpha^* \frac{\sigma}{\sqrt{n}},\ \ \bar{X} + z_\alpha^* \frac{\sigma}{\sqrt{n}}\ \right]$$

数式表現のまとめ（2/2）

$$-z_\alpha^* \le \frac{\mu - \bar{X}}{\sqrt{\sigma^2/n}} \le z_\alpha^*$$

Z は 標準正規分布 $N(0,1)$ に従う

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



両側 α 点

$$P\left\{\left|\frac{\mu - \bar{X}}{\sigma/\sqrt{n}}\right| \le z_\alpha^*\right\} = 1-\alpha$$

場合２

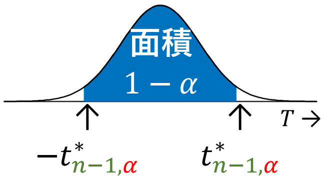
	母集団 全ての個体、正規分布		標本 無作為に抽出した n 個	
	母平均 μ	母分散	標本平均 $\bar{X}$	標本不偏分散 $\hat{\sigma}^2$
場合2	推定する	未知	データから計算	

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$-t_{n-1,\alpha}^* \le \frac{\mu - \bar{X}}{\hat{\sigma}/\sqrt{n}} \le t_{n-1,\alpha}^*$$

↑

自由度 n-1 の **t 分布** に従う



$t_{n-1,\alpha}^*$ は両側 α 点
信頼度 $100(1-\alpha)$ %

$t_{9,\,0.05}^* = 2.262$ $(n=10)$
$t_{19,0.05}^* = 2.093$ $(n=20)$

母平均の信頼区間（まとめ）

$100(1-\alpha)\ \%$

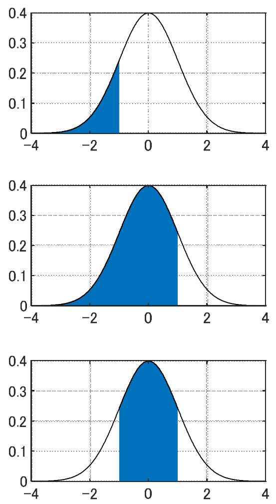
場合１（母分散が既知）

$$\mu \in \left[\ \bar{X} - z_{\alpha}^{*}\frac{\sigma}{\sqrt{n}},\quad \bar{X} + z_{\alpha}^{*}\frac{\sigma}{\sqrt{n}}\ \right]$$

場合２（母分散が未知）

$$\mu \in \left[\ \bar{X} - t_{n-1,\alpha}^{*}\frac{\hat{\sigma}}{\sqrt{n}},\quad \bar{X} + t_{n-1,\alpha}^{*}\frac{\hat{\sigma}}{\sqrt{n}}\ \right]$$

Rで区間推定 pnorm



標準正規分布の
青い部分の面積

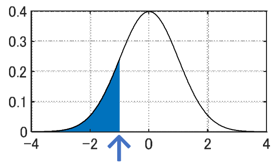
```
> pnorm(-1, mean=0, sd=1)
[1] 0.1586553

> pnorm(+1, mean=0, sd=1)
[1] 0.8413447

> pnorm(+1)-pnorm(-1 )
[1] 0.6826895
```

mean=0, sd=1
のときは省略可

Rで区間推定 qnorm



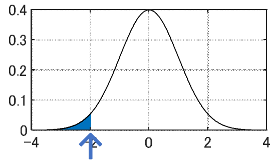
片側α 点

青い部分の面積が指定した値となるときの横軸の値

```
> qnorm(0.1586553)
[1] -0.9999998

> qnorm(pnorm(-1))
[1] -1

> alp <- 0.05
> qnorm(alp/2)
[1] -1.959964
```



片側α 点

両側α 点 z_{α}^{*} は qnorm(α/2) で計算

場合１ ネジの長さは同じか？

今までは、平均 19 mm でネジを生産していた。
標準偏差は 2 mm だった。

機械を刷新したら、長くなったようなので、
調べて欲しい。

機械を刷新した後（無作為に抽出したサンプル）
ネジの長さは、 {20, 21, 21, 19, 24}

母平均 μ の信頼区間 95% を示す
ここで、母標準偏差 $\sigma = 2$ とする

場合１

Ｒのプログラム

データを準備
標本の数 $n = 5$
標本平均 $\bar{X} = 21$
母標準偏差 $\sigma = 2$

```
x   <- c(20,21,21,19,24)
n   <- length(x)
xbr <- mean(x)
sgm <- 2
```

両側α点 $z_\alpha^* = 1.96$
（α =0.05 で 95%）

$\Delta = z_\alpha^* \frac{\sigma}{\sqrt{n}} = 1.75$

```
alp <- 0.05
z   <- -qnorm(alp/2)
dlt <- z*sgm/sqrt(n)
```

95%の信頼区間

$\mu \in [\bar{X} - \Delta,\ \ \bar{X} + \Delta]$
$= [19.2, 22.8]$

```
c(xbr-dlt,xbr+dlt)
[1] 19.24695 22.75305
```

今までは 19 だったが、刷新後は長くなった

場合２ ネジの長さは同じか？

今までは、平均 19 mm でネジを生産していた。
なお、標準偏差は測定していないので不明。

機械を刷新したら、長くなったようなので、
調べて欲しい。

機械を刷新した後（無作為に抽出したサンプル）
ネジの長さは、｛20, 21, 21, 19, 24｝

↓

母平均 μ の信頼区間 95% を示す
この場合、**標本標準偏差** $\hat{\sigma}$ を使う

場合２

Ｒのプログラム

データを準備

標本の数　$n = \ 5$

標本平均　$\bar{X} = 19 + \mathbf{2}$

標本標準偏差 $\hat{\sigma} = 1.87$

両側α点 $t^*_{n-1,\alpha} = 2.78$

（α =0.05 で 95%）

$\Delta = t^*_{n-1,\alpha}\frac{\hat{\sigma}}{\sqrt{n}} = 2.32$

95%の信頼区間

$\mu \in [\bar{X} - \Delta, \quad \bar{X} + \Delta]$

$= [18.7, 23.3]$

```
x    <- c(20,21,21,19,24)
n    <- length(x)
xbr  <- mean(x)
sgm  <- sd(x)

alp  <- 0.05
z    <- -qt(alp/2,n-1)
dlt  <- z*sgm/sqrt(n)

c(xbr-dlt,xbr+dlt)
[1] 18.67706 23.32294
```

今までは19で、刷新後の ＋２ はバラツキの範囲内

場合２別解

Ｒの特長を活かした超簡単なプログラム

たったこれだけ

```
> x    <- c(20,21,21,19,24)
> t.test(x,mu=19)
```

```
        One Sample t-test
data:  x
t = 2.3905, df = 4, p-value = 0.07513
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 18.67706 23.32294
sample estimates:
mean of x
       21
```

← ｐ値

95%の信頼区間を一瞬で計算

$\mu \in [18.7, 23.3]$

まとめ

- 今までは、長さが平均で 19 となる様に生産していた
- 機械を刷新したら、長さが変わったようだ
- 5個を抽出して測ったら｛20, 21, 21, 19, 24｝だった
- これらの平均は 21 なので、今までとの「差は +2」
- これでは、規格を満たさないのでは？

長さの平均は無作為に5個を抽出する度にバラつくが、

場合１（母分散が **既**知）では、規格の基準19 が、信頼区間である[19.2, 22.8] の範囲**外** なので、長さの差が無い とは言えない

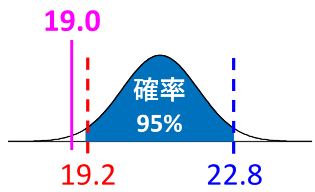
場合２（母分散が **未**知）では、規格の基準19 が、信頼区間である[18.7, 23.3] の範囲**内** なので、長さの差が無い と言える

帰無仮説

場合１ では、規格の基準19 が、信頼区間である[19.2, 22.8] の範囲**外** なので、

長さには差がありそう
（帰無仮説は 棄却される）

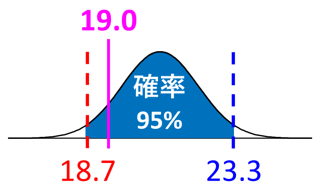
※ 長さの差は「有意」である。



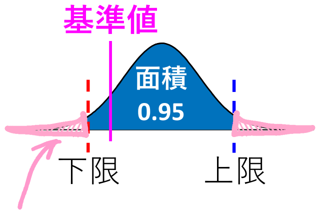
場合２ では、規格の基準19 が、信頼区間である[18.7, 23.3] の範囲**内** なので、

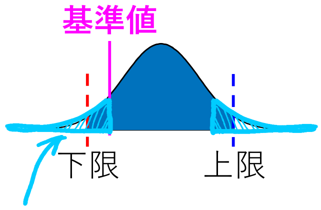
長さには差が無さそう
（帰無仮説は 採択される）

※ nが小さすぎるかも。長さが同じとまでは言えない。



「p 値」



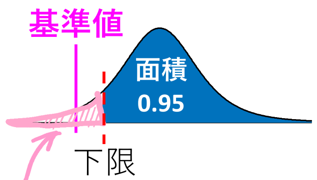


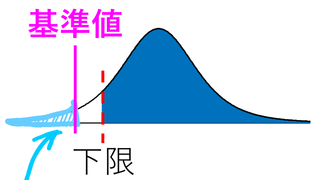
この部分の面積は 両側で
1-0.950
= 0.050 ←

この部分の面積が
「p 値」
→ 0.0??

どちらが
大きいかで判断する

「片側検定」





この部分の面積は
1-0.950
= 0.050 ←

この部分の面積が
「p 値」
→ 0.0??

どちらが
大きいかで判断する

2標本問題

場合３

	母集団×2つ 全ての個体、正規分布		標本×2つ 無作為抽出した m 個と n 個	
	母平均の差 $\mu_1-\mu_2$	母分散 σ_1^2 と σ_2^2	標本平均 $\bar{X}_1$ と $\bar{X}_2$	標本分散
場合3	推定する	既知	データから計算	不要

$\bar{X}_1=\frac{1}{m}\sum_{i=1}^{m}X_{1,i}$　$\bar{X}_2=\frac{1}{n}\sum_{i=1}^{n}X_{2,i}$

母平均の差 $\mu_1-\mu_2$ の$100(1-\alpha)$%の信頼区間

$$\mu_1-\mu_2\in\left[\bar{X}_1-\bar{X}_2-z_\alpha^*\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}},\ \ \bar{X}_1-\bar{X}_2+z_\alpha^*\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right]$$

場合４

	母集団×2つ 全ての個体、正規分布		標本×2つ 無作為抽出した m 個と n 個	
	母平均の差 $\mu_1 - \mu_2$	母分散 σ_1^2 と σ_2^2	標本平均 $\bar{X}_1$ と $\bar{X}_2$	合併 標本分散 $\hat{\sigma}^2$
場合3	推定する	未知	データから計算	

$$\hat{\sigma}^2 = \frac{1}{m+n-2}\left\{\sum_{i=1}^{m}\left(X_{1,i} - \bar{X}_1\right)^2 + \sum_{i=1}^{n}\left(X_{2,i} - \bar{X}_2\right)^2\right\}$$

母平均の差 $\mu_1 - \mu_2$ の$100(1-\alpha)$ %の信頼区間

$$\mu_1 - \mu_2 \in \left[\bar{X}_1 - \bar{X}_2 - t^*_{m+n-2,\alpha}\sqrt{\frac{\hat{\sigma}^2}{m} + \frac{\hat{\sigma}^2}{n}},\ \bar{X}_1 - \bar{X}_2 + t^*_{m+n-2,\alpha}\sqrt{\frac{\hat{\sigma}^2}{m} + \frac{\hat{\sigma}^2}{n}}\right]$$

付録１

Ｒをインストールする

Rをダウンロード（1/2）

https://cran.r-project.org/

The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Search

About R
R Homepage
The R Journal

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

base — Binaries for base distribution. This is what you want to **install R for the first time**.

Rをダウンロード（2/2）

R-4.2.1 for Windows

CRAN
Mirrors
What's new?
Search

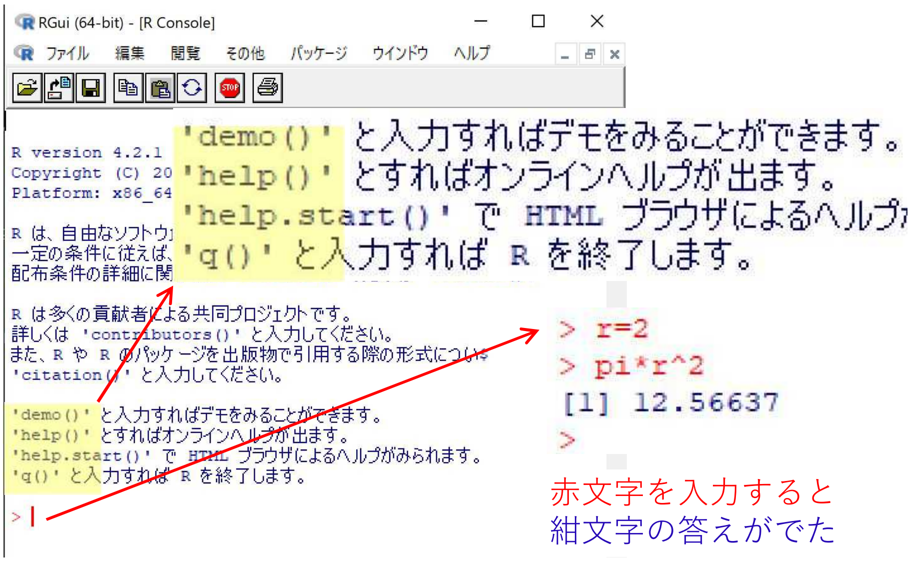
Download R-4.2.1 for Windows (79 megabytes, 64 bit)
README on the Windows binary distribution
New features in this version

This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT has to be installed manually from here.

R-4.2.1-win.exe（78.7MB）を入手して実行する

このアイコンをクリック
するとＲが動く

Rを動かしてみる



赤文字を入力すると
紺文字の答えがでた

Rで計算してみる

```
> r=2
> pi*r^2
[1] 12.56637
```

☜ R に入力：赤文字
☜ R の出力：紺文字

$r = 2$
πr^2

記号「<-」は代入

```
> r <- c(2,3,4)
> pi*r^2
[1] 12.56637 28.27433 50.26548
```

$\boldsymbol{r} = [2 \quad 3 \quad 4]$
$\pi \boldsymbol{r}^2$

```
> pi*r^2 -> a
> mean(a)
[1] 30.36873
```

```
> sum(a)/3
[1] 30.36873
```

$\boldsymbol{a} = \pi \boldsymbol{r}^2$
$E[\boldsymbol{a}]$

統計量を計算してみる

データ	x <- c(1, 2 ,3, 4)	x_i , i=1,2,3, ・・・,n
平均	mean(x) 2.5	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
不偏分散	var(x) 1.666667	$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$
標準偏差	sd(x) 1.290994	$\hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}$

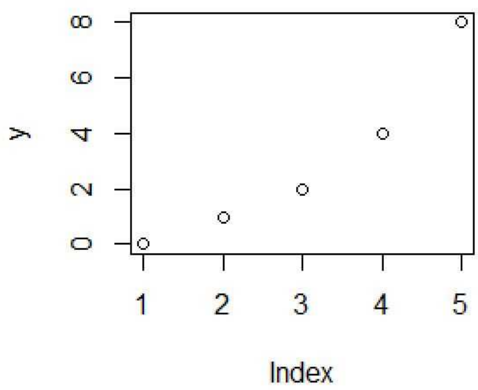
プロットしてみる（1/3）

データ y をプロットしたい

$\boldsymbol{y} = [0 \quad 1 \quad 2 \quad 4 \quad 8]$

```
y <- c(0,1,2,4,8)
plot(y)
```

R に入力



プロットしてみる（2/3）

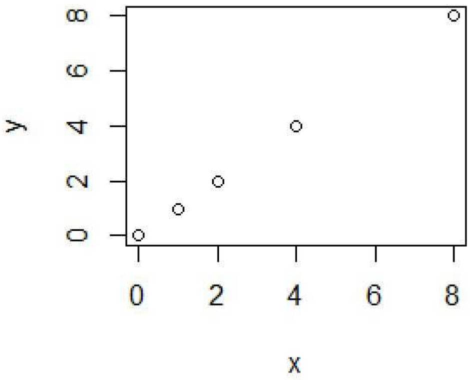
x と y の**散布図**をみたい

$\boldsymbol{x} = [0 \quad 1 \quad 2 \quad 4 \quad 8]$

$\boldsymbol{y} = [0 \quad 1 \quad 2 \quad 4 \quad 8]$

```
x <- c(0,1,2,4,8)
y <- c(0,1,2,4,8)
plot(x,y)
```

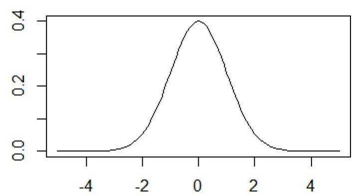
R に入力



プロットしてみる（3/3）

-5 から 5 までの範囲の
標準正規分布

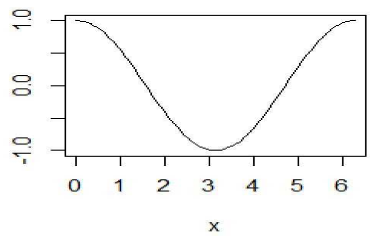
```
curve(dnorm, -5,5)
```



0 から 2π までの範囲で
cos をプロット

```
plot(cos, 0, 2*pi)
```

R に入力



付録２

カイ自乗分布　と　ティー分布

各種統計量の確率分布

全データ

母平均　μ
母分散　σ^2

正規分布

無作為に抽出

標本のデータ

標本の平均

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

標本の不偏分散

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

$\bar{X}$は正規分布
（平均 μ, 分散 σ^2/n）

$\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ は
平均 0 , 分散 1 の正規分布

$\dfrac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$ は
自由度 $n-1$ のティー分布

$\dfrac{(n-1)\hat{\sigma}^2}{\sigma^2}$ は
自由度 $n-1$ のカイ自乗分布

各種統計量の区間推定

$1-\alpha$ の確率で
以下の区間に入る

$\bar{X}$は正規分布に従う
（平均 μ, 分散 σ^2/n）

$$\mu \in \left[\bar{X} - z_\alpha^* \frac{\sigma}{\sqrt{n}},\ \ \bar{X} + z_\alpha^* \frac{\sigma}{\sqrt{n}}\right]$$

⬅ $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ は
平均 0 , 分散 1 の正規分布

$$\mu \in \left[\bar{X} - t_{n-1,\alpha}^* \frac{\hat{\sigma}}{\sqrt{n}},\ \ \bar{X} + t_{n-1,\alpha}^* \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

⬅ $\dfrac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$ は
自由度 $n-1$ のティー分布

$$\sigma^2 \in \left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1,\frac{\alpha}{2}}},\ \ \frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]$$

⬅ $\dfrac{(n-1)\hat{\sigma}^2}{\sigma^2}$ は
自由度 $n-1$ のカイ自乗分布

正規分布から誘導される分布

$$Z_1, Z_2, \cdots, Z_n \sim N(0,1)$$

⬅ Z_i は標準正規分布に
従う独立な確率変数

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ のとき}$$
$$Y \sim \chi_n^2$$

⬅ Y は自由度 n の
カイ自乗分布に従う

$$X \sim N(0,1),\ Y \sim \chi_n^2 \text{ のとき}$$
$$T = \frac{X}{\sqrt{Y/n}} \sim t_n$$

⬅ T は自由度 n の
ティー分布に従う

様々な分布と密度関数

正規分布 $X \sim N(\mu,\sigma)$ ➡ $f(x|\mu,\sigma) = \dfrac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$

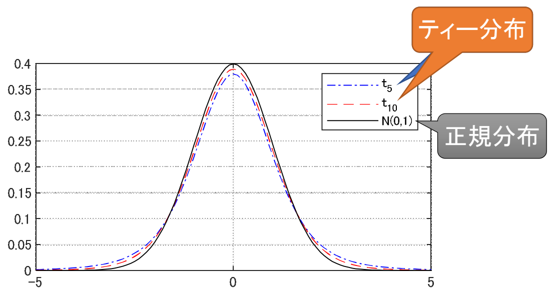
自由度 n の
カイ自乗分布 $X \sim \chi_n^2$ ➡ $f(x|n) = \dfrac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} exp\left(-\dfrac{x}{2}\right)$

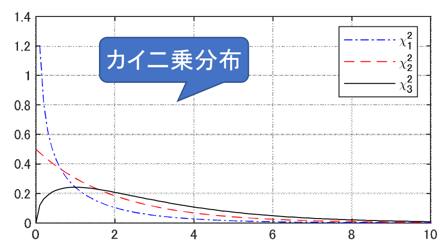
$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1}\, dt$ （ガンマ関数）

自由度 n の
ティー分布 $X \sim t_n = \dfrac{N(0,1)}{\sqrt{\chi_n^2/n}}$ ➡ $f(x|n) = \dfrac{1}{\sqrt{n}\,B\left(\frac{n}{2},\frac{1}{2}\right)}\left(1+\dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$

$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$ （ベータ関数）

様々な分布と密度関数






```
% MATLAB Program

close all; clear all;

%--------------------
% t 分布
%--------------------
x =-5:.1:5;
y1 = tpdf(x,5);
y2 = tpdf(x,10);
y3 = normpdf(x,0,1);

figure('Position',[10 10 400 200]);
plot(x,y1,'B-.'); hold on
plot(x,y2,'R--'); hold on;
plot(x,y3,'K-')
legend({'t_5','t_{10}','N(0,1)'})
grid on;

%--------------------
% カイ二乗分布
%--------------------
x = 0:.1:10;
y1 = chi2pdf(x,1);
y2 = chi2pdf(x,2);
y3 = chi2pdf(x,3);

figure('Position',[10 210 400 200]);
plot(x,y1,'B-.'); hold on
plot(x,y2,'R--'); hold on;
plot(x,y3,'K-')
legend({'¥chi^2_1','¥chi^2_2','¥chi^2_3'})
grid on;
```

初版：　2022年8月

制作：　岩橋政宏
所属：　長岡技術科学大学