

確率・統計

第1回 イントロ

兵庫県立大学 社会情報科学部

川嶋宏彰

kawashima@sis.u-hyogo.ac.jp

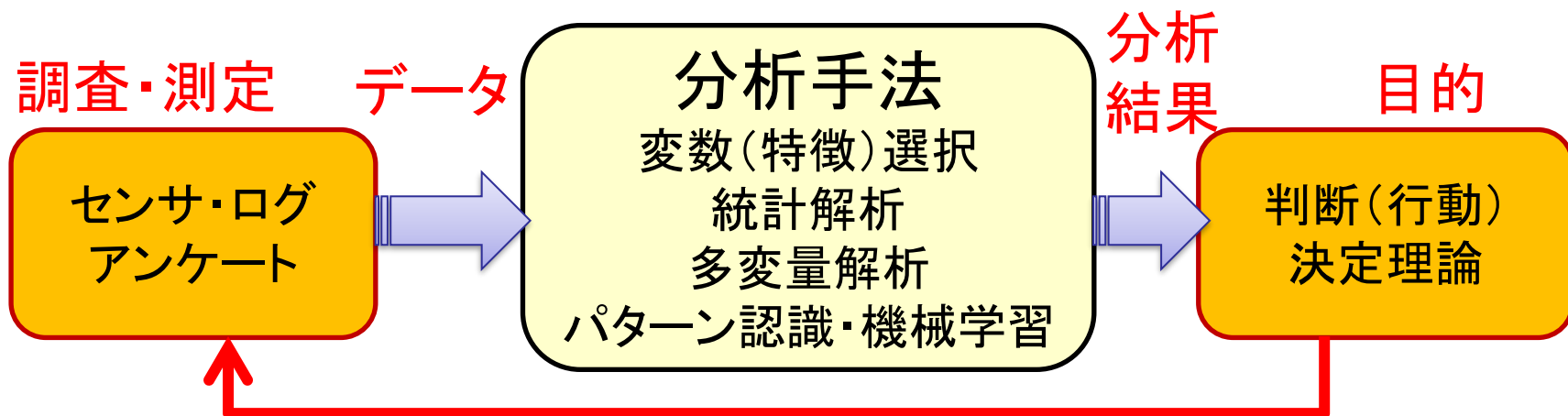
- なぜ統計学を学ぶのか？
 - 統計学と機械学習
 - 統計学の歴史（覚えておきたい名前）
- アンケート
- 「確率・統計」では何を学ぶのか？
 - この講義では確率論よりも統計学をより深く学ぶ
 - 前期・統計学との差分は？
- Rコマンドの使い方
 - 統計解析のツールとして
 - Rを学ぶ上でも参考になる

データ分析の流れ

3

- データと目的に応じて適切な分析を行う

1. 適切にデータ収集（調査，実験・試験，ロギング）
2. 分析手法
3. 分析結果に基づく判断（何を明らかにしたいのか？）
4. 判断に基づく行動・改善



統計学と機械学習

4

- 統計学と機械学習はオーバーラップがある

統計学

確率・統計

- 記述統計
- 推測統計
 - 仮説検定
 - カイ二乗検定
 - t検定
 - 分散分析
 - ：
 - 統計的推定
 - 点推定
 - 区間推定
 - ：
 - 回帰・相関分析

多変量解析

- 多変量確率分布
- 数量化I, II, ..., 類
- 重回帰分析

データマイニング・

人工知能・機械学習

- 教師あり学習
 - 回帰
 - 分類
- 教師なし学習
 - クラスタリング
 - 潜在変数モデル
 - 異常検知
- 強化学習

- 統計学は人へ判断材料を提供する（ことが多い）
 - 記述統計，推測統計（区間推定，仮説検定）
- 機械学習は予測精度を第一目標とする（ことが多い）
 - 人を介さない自動判断も範疇（例：自動運転，株の自動取引）
 - 必ずしも判断の過程が見えなくてもよい（例：顔認識，物体認識）
 - ただし説明可能性 (explainability) ・ 解釈可能性 (interpretability) もしばしば関心事（特に今とても盛り上がっている）
- <https://towardsdatascience.com/predicting-vs-explaining-69b516f90796>

- 同じ手法でも見るポイントが異なる（例：回帰分析）
 - 統計学：結果を予測するためにどの要因が重要か
 - 機械学習：状況が多少変わっても予測精度は落ちないか
- 「統計学と機械学習は何が違うのか？」はよく質問される
 - 人によって答えは違う．そもそもどこまで「機械学習」とするか？
 - オーバラップも多い
 - いくつかの記事を見ていろいろな意見を知ることをお勧め
 - 「わかる」vs.「できる」？
 - 統計学はどこまでを保証するのか？
- とはいえ，まずは双方学んでみなければよくわからない

記述統計と推測統計

7

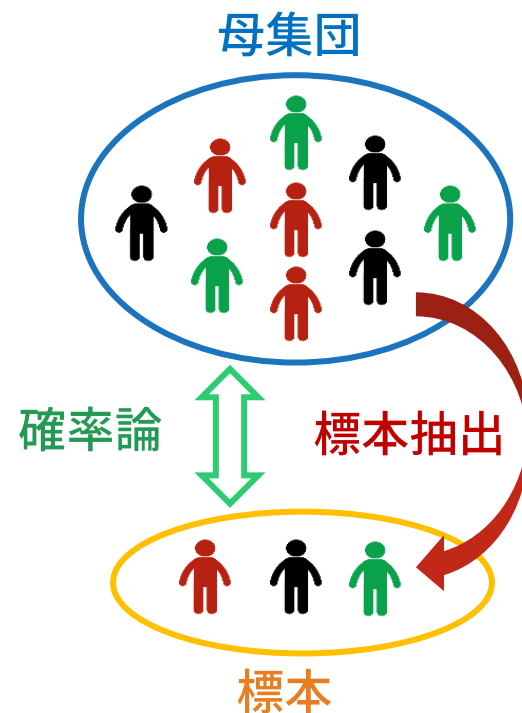
- 統計には「記述統計」と「推測統計」がある

- 記述統計 (descriptive statistics)

- データを要約・視覚化することで理解
- 平均, 中央値, 分散, 標準偏差, 相関係数, ヒストグラム, 散布図

- 推測統計 (inferential statistics) (推計統計)

- サンプリング調査を前提 (右図)
- 部分から全体を知る
- 仮説が正しいかを判断する (仮説検定)
- 母集団の統計量を見積もる (推定)
- 過去から未来を予測する



推測統計も前期「統計学」で一部学んだ

探索的か確認的か

- 仮説検証的データ分析の前にデータをよく見ること
 - 探索的データ分析 (exploratory data analysis)
 - まずはデータの特徴をよく知る
 - 各種統計量の確認，可視化（散布図，ヒストグラム・・・）などを通じて解析対象や仮説を明確化していく
 - 主に記述統計
 - 確認的(検証的)データ分析 (confirmatory data analysis)
 - 仮説を検証する
 - 推測統計（標本調査など） ・ 記述統計

- 記述統計から推測統計へ

- ～18世紀：統計の誕生
 - 国家 (Staat) の実態をとらえるための統計
 - 今日の国勢調査などに対応
 - ラプラス(数学者)：古典確率論の成立
- 19世紀：記述統計の成立
 - ガウスによる誤差や正規分布の研究 → 社会科学へも影響
 - 「近代統計学の父」 ケトレー：社会物理学
 - 「平均人」の概念・身体的データ (Body Mass Index: BMI)
 - ゴルトン (ダーウィンの従弟)：進化論 → 生物統計 / 相関や回帰
 - カール・ピアソン：生物統計から数理統計学 / 記述統計を大成

- 記述統計から推測統計へ

- 20世紀：推測統計の成立

- ゴセット：ビールの醸造の研究からStudent's t 分布を発見
 - フィッシャー：推測統計学の確立，農事試験場の実験計画研究
 - ネイマン / エゴン・ピアソン：推測統計学の確立
 - 信頼区間・仮説検定などの理論的体系を構築
 - (ベイズ統計：主観的な確率)

- 21世紀：あらゆる分野で活用

- 経済・物理・疫学・生物・インターネット・マーケティング
 - コンピュータの利用

- なぜ統計学を学ぶのか？
 - 統計学と機械学習
 - 統計学の歴史（覚えておきたい名前）
- アンケート
- 「確率・統計」では何を学ぶのか？
 - この講義では確率論よりも統計学をより深く学ぶ
 - 前期・統計学との差分は？
- Rコマンドの使い方
 - 統計解析のツールとして
 - Rを学ぶ上でも参考になる

アンケート

12

- 一部抜粋し講義で使用予定（個人情報を入れないように）



<https://forms.gle/WSJhaEUmvZdUMrJM8>

- なぜ統計学を学ぶのか？
 - 統計学と機械学習
 - 統計学の歴史（覚えておきたい名前）
- アンケート
- 「確率・統計」では何を学ぶのか？
 - この講義では確率論よりも統計学をより深く学ぶ
 - 前期・統計学との差分は？
- Rコマンドの使い方
 - 統計解析のツールとして
 - Rを学ぶ上でも参考になる

データの種類

14

- データは質的データと量的データに分けられる

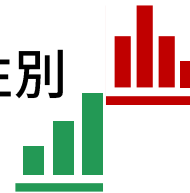
- 量的データ：数値を値としてとる

- 連続尺度：時間, 50kg, 5mmHg, 30歳
(間隔尺度や比例尺度など)



- 質的データ：記号を値としてとる

- 名義尺度（例）血液型, 投与有無, 性別
- 順序尺度（例）サイズ {S, M, L}



各行が各個人のデータ

年齢	学歴	教育年数	結婚	職業	家族関係	人種	性別	週の労働時間	母国	収入 \$
42	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	40	United-States	>50K
37	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	80	United-States	>50K
30	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	40	India	>50K
23	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	30	United-States	<=50K
32	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	50	United-States	<=50K
40	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	40	?	>50K
34	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	45	Mexico	<=50K
25	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	35	United-States	<=50K

Census income データセットより

統計学における要因と結果

15

- 要因と結果の関係を知りたい（予測につながる）

→ 判断・意思決定の材料



連続尺度



名義尺度



順序尺度

- 新薬は血圧に対して効果があるか？

• 要因：{新薬のむ or 新薬のまない} → 結果：血圧が?mmHg下がった

- 精密検査にまわす必要はあるか？

• 要因：空腹時血糖値? mg/dl, 食後血糖値? mg/dl

→ 結果：{検査必要 or 検査不要}





- お客さんに商品Aを薦めると購買行動につながるか？

• 要因：{男性 or 女性}, 年齢?才, {過去に購入あり or 過去に購入なし}
→ 結果：{商品Aを購入する or 商品Aを購入しない}

様々なデータ分析手法

16

- データや分析目的に合わせ適切なデータ分析手法を選ぶ
 - データの種類，サイズ，収集条件などを考慮


		要因・条件（説明変数）	
		量的 	質的 
結果（目的変数）	量的 	(散布図) 回帰分析 ニューラルネット サポートベクター回帰	(箱ひげ図) t検定 分散分析 数量化I類（回帰分析+ダミー変数） 回帰木
	質的 	ロジスティック回帰分析 判別分析 ニューラルネット サポートベクターマシン ナイーブベイズ分類器	(クロス表，分割表) カイ二乗検定 フィッシャーの正確検定 数量化II類（判別分析+ダミー変数） 決定木

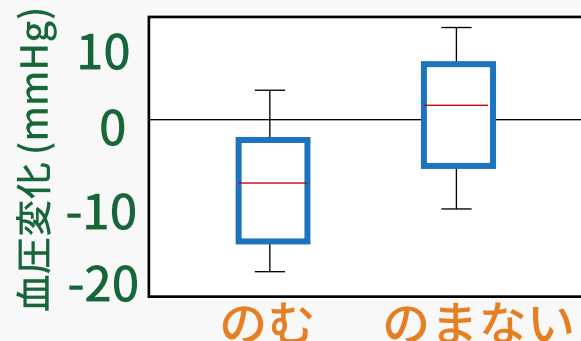
新薬は血圧に対して効果があるか？





要因：{新薬のむ or 新薬のまない}

 名義尺度

→ 結果：血圧が 0mmHg 下がった

 連続尺度



		要因・条件 (説明変数)	
		量的 	質的 
結果 (目的変数)	量的 	(散布図) 回帰分析 ニューラルネット サポートベクター回帰	(箱ひげ図) t検定 分散分析 数量化I類 (回帰分析+ダミー変数) 回帰木
	質的 	ロジスティック回帰分析 判別分析 ニューラルネット サポートベクターマシン ナイーブベイズ分類器	(クロス表, 分割表) カイ二乗検定 フィッシャーの正確検定 数量化II類 (判別分析+ダミー変数) 決定木

平均に違いはあるのか？

18

• 2つのグループの平均に差があるか調べる

- (例) 食○ログによればラーメンA店は平均3.1点，B店は平均3.3点
- 兵子さんは「ラーメン激戦区の老舗A店も，○○系ブームに押されB店より評価が落ちている」と結論づけたがこれは正しいのか？
(ラーメン店A, Bを評価した人は全員同程度の基準を持つと仮定)



A店 (38評価)	
評価ID	点数
1	2
2	3
3	1
4	2
...	...
38	3

平均 3.1

B店 (41評価)	
評価ID	点数
1	3
2	2
3	5
4	4
...	...
41	4

平均3.3



仮説検定
(t検定)

出口調査と得票率

19

- 出口調査から得票率の信頼区間を推定する（区間推定）
 - 候補者A, Bの選挙で，出口調査ではAが63票，Bが37票であった
 - Aの最終的な得票率はどの範囲で見積もればよいか？

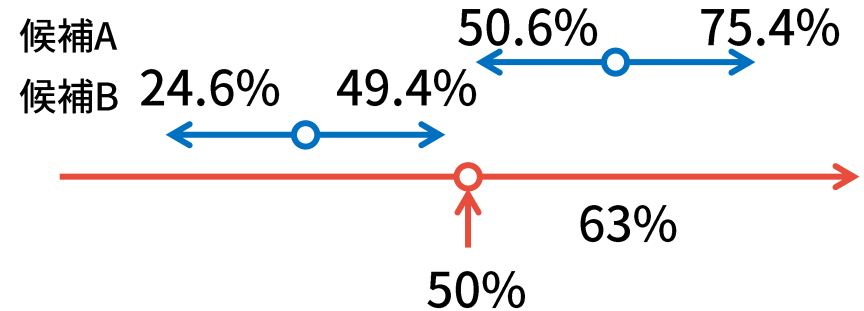
出口調査データ

対象者ID	投票先
1	B
2	B
3	A
4	B
...	...
100	A



集計表

候補A	候補B	計
63	37	100



母比率の区間推定
○○%信頼区間
(上の例は99%)

分布に偽りはないか？

20

- 適合度検定で想定と異なる分布を見分ける
 - (例) 子供会のくじ屋に「1等: 10%, 2等: 30%, 外れ: 60%」の看板
 - 神太君は1万円つき込んで100回くじを引きログを取った
 - 神太君は文句を言ったが、くじ屋のおじさんには偶然といわれた
 - 再度文句を言うべきか？それとも偶然なのか？

整形済みログデータ

試行	0 (外れ), 1, 2
1	0
2	2
3	0
4	1
...	...
100	0



集計表

1等	2等	外れ	計
2	20	78	100

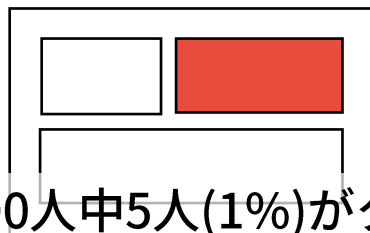


適合度検定
(カイ二乗検定)

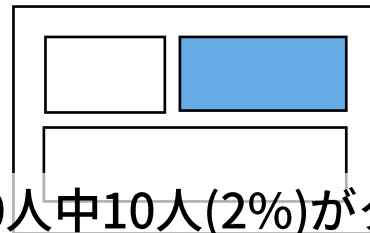
ウェブのA/Bテスト

21

- A/Bテストでどちらのデザインがよいかを調べる
 - WebサービスやWeb広告のデザイン
 - メニュー配置，色，広告の画像・・・の最適化 (optimization)
 - ユーザ訪問時に出すデザインを変える → どちらがクリック率が高いか？



A: 500人中5人(1%)がクリック



B: 500人中10人(2%)がクリック

整形済みログデータ

ユーザ番号	A or B	クリック？
1	A	1
2	B	0
...
1000	B	1



クロス集計表（分割表）

	No	Yes
A	495	5
B	490	10



仮説検定
(母比率の検定,
カイ二乗検定)

「確率・統計」の範囲（シラバスより）

22

1. 確率・統計の概要
2. データの要約
3. 確率と確率分布
4. 確率論と統計学
5. 大数の法則と中心極限定理
6. 母平均の検定と推定
7. 分散分析
8. 中間テストとこれまでのまとめ
9. データ収集と実験計画
10. 母比率の検定と推定
11. カイ二乗検定
12. サンプルサイズと検定
13. 相関と回帰
14. 回帰分析
15. まとめと発展的话题
16. 評価（到達度の確認）

平均の検定？

23

- 前期・統計学より
 - ある工場では毎日500gの製品を多数製造している。
 - 50個を無作為に抽出して重さを量ったところ、平均495gであった。
 - このことから、この工場の製品の重さの平均値は500gではないと判断してよいだろうか。
 - この工場の製品の重さは標準偏差16gの正規分布に従うと仮定し、有意水準5%で仮説検定を行え。
- 前期は「母分散が既知の場合の、1群の母平均の検定」まで

「確率・統計」の範囲（シラバスより）

24

前期の内容中心

1. 確率・統計の概要
2. データの要約
3. 確率と確率分布
4. 確率論と統計学
5. 大数の法則と中心極限定理
6. 母平均の検定と推定
7. 分散分析
8. 中間テストとこれまでのまとめ
9. データ収集と実験計画
10. 母比率の検定と推定
11. カイ二乗検定
12. サンプルサイズと検定
13. 相関と回帰
14. 回帰分析
15. まとめと発展的话题
16. 評価（到達度の確認）

前期・統計学の内容を含みつつ、様々な推測統計を紹介

「統計検定® 2 級」相当の内容

レポート・中間テスト60%
定期試験40%を基準

- 前期・統計学の内容を含みつつ，様々な推測統計を紹介
 - 2群の平均や比率の差の検定（2標本検定），3群以上の場合も扱う
 - 標本サイズが小さいとき
 - 標本平均の検定で「母分散が未知」の場合も扱う
 - 相関の検定，回帰係数の検定，重回帰，ロジスティック回帰も含む

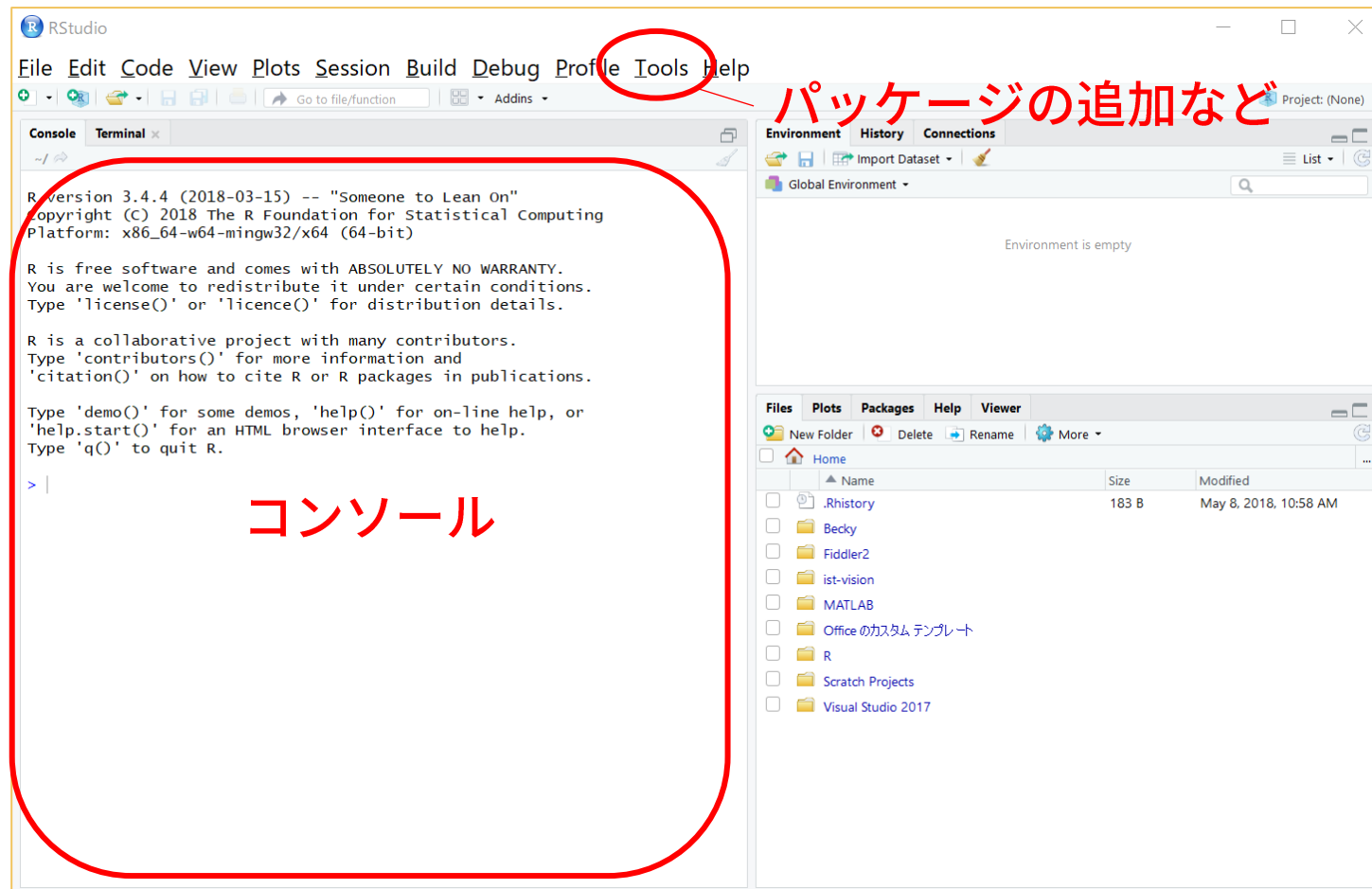
- なぜ統計学を学ぶのか？
 - 統計学と機械学習
 - 統計学の歴史（覚えておきたい名前）
- アンケート
- 「確率・統計」では何を学ぶのか？
 - この講義では確率論よりも統計学をより深く学ぶ
 - 前期・統計学との差分は？
- Rコマンドの使い方
 - 統計解析のツールとして
 - Rを学ぶ上でも参考になる

- プログラミング言語「R」の機能をプログラミング無しで使うことのできる統計解析ツール
 - 実行時に「R」のスクリプトも出力されるため、「R」のプログラミングを学ぶ上でも参考になる
 - 出力されたスクリプトを変更して実行することもできる

さっそくRコマンダーを使ってみよう

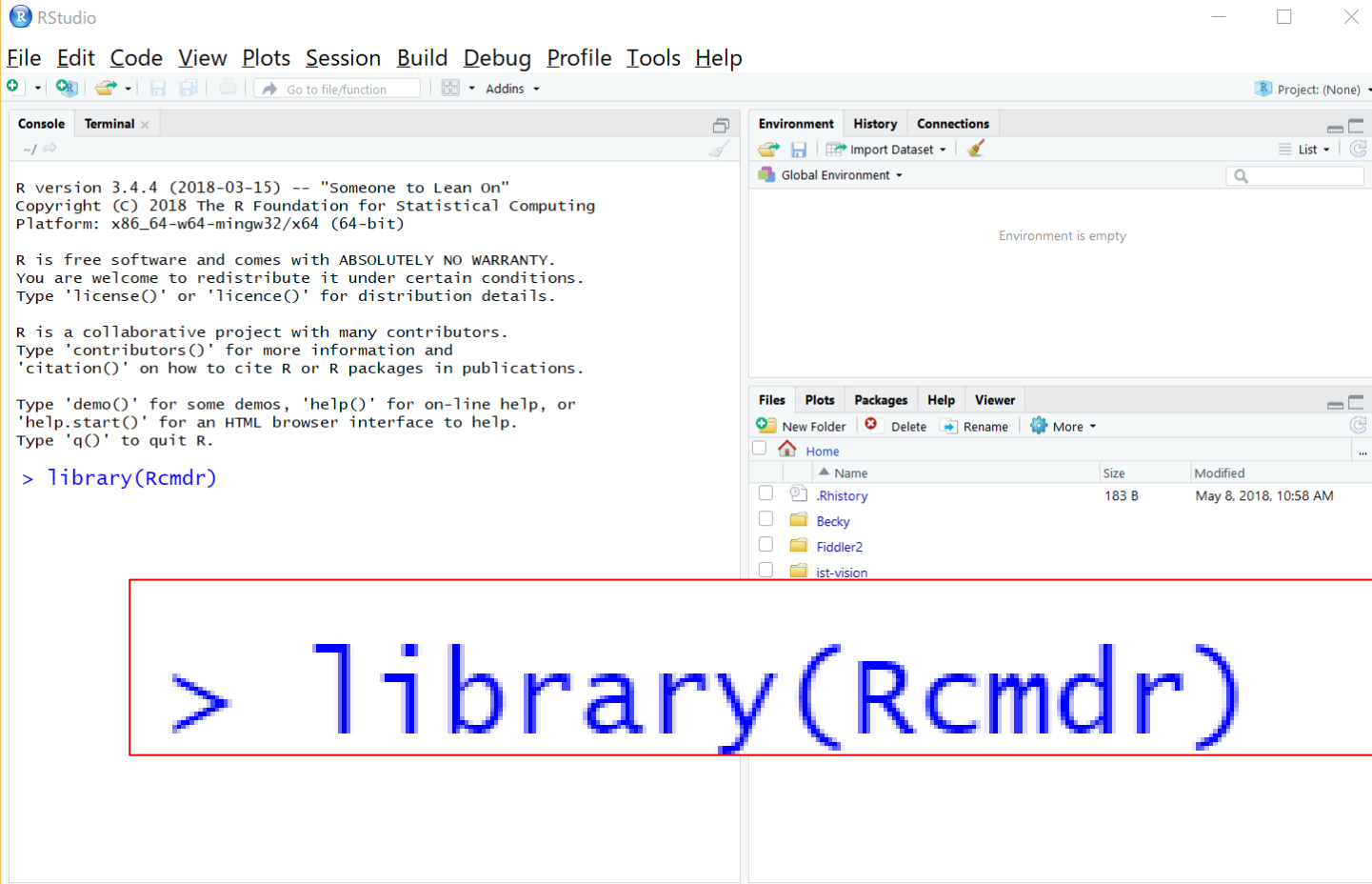
28

- R, R Studio, Rコマンダー
 - R (R Console)
 - Rのスクリプトを使うための基本ツール
 - 単体でもRが使えるがRstudioだとより使いやすい
 - RStudio
 - Rの統合開発環境 (IDE). R Console, エディタや履歴, ファイルエクスプローラなど, さまざまな支援ツールが統合されている
 - Rコマンダーは, Rの一つの「パッケージ」でRがインストールされている必要あり
- Rもしくは RStudio のコンソールで以下を実行すると起動
 - > library(Rcmdr)



Rコマンドーの起動

30



The screenshot shows the RStudio application window. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar. The main workspace is divided into three panes: Console, Environment, and Files. The Console pane shows the R startup messages and the command `> library(Rcmdr)` entered. The Environment pane shows the Global Environment, which is empty. The Files pane shows the Home directory with a list of files and folders.

```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

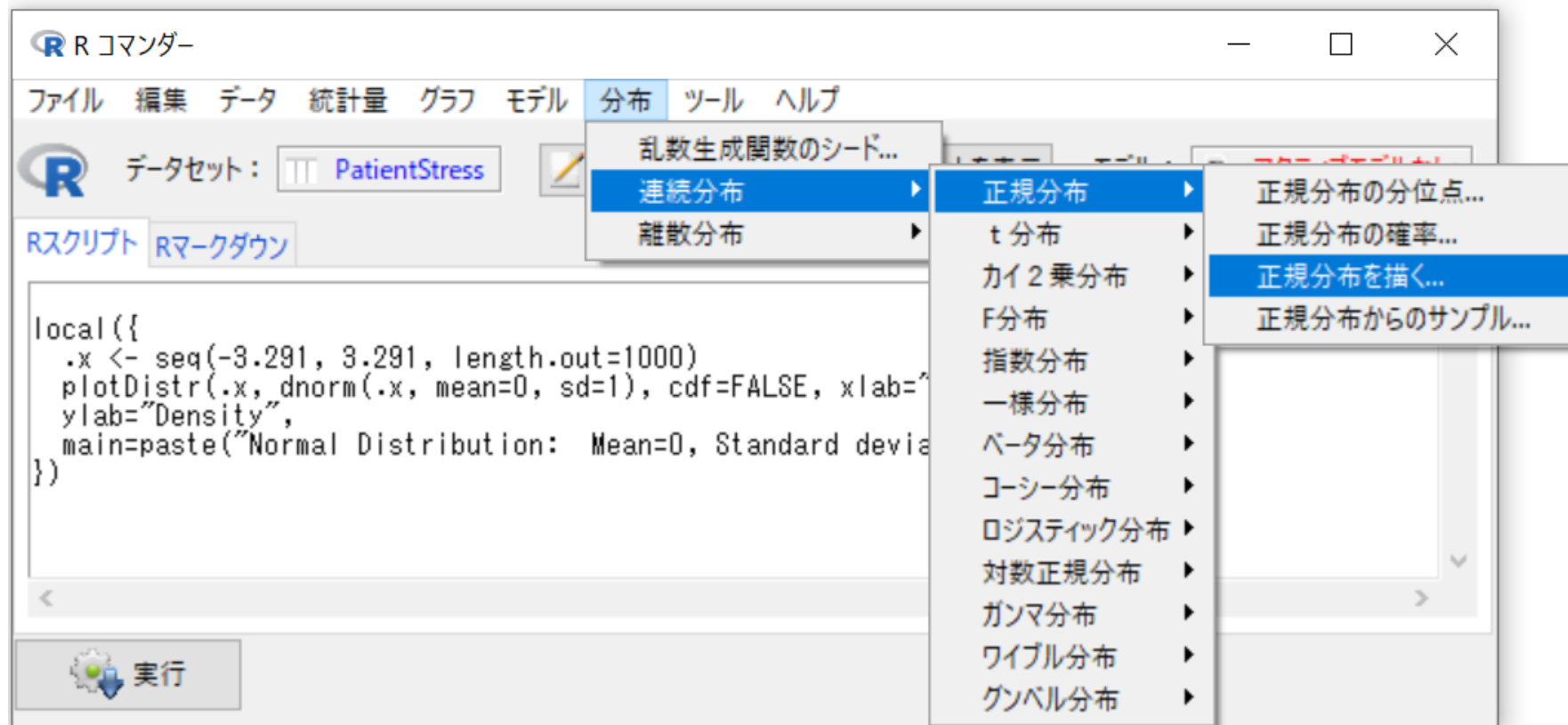
> library(Rcmdr)
```

`> library(Rcmdr)`

正規分布を表示してみよう

31

- 分布 > 連続分布 > 正規分布 > 正規分布を描く

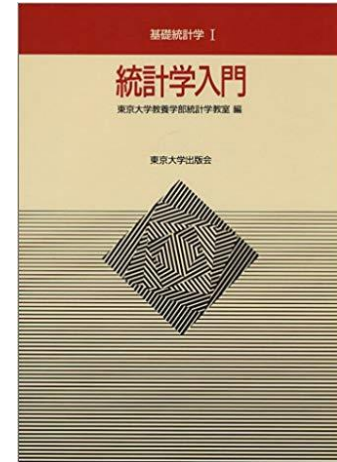


- 教科書

- 東大出版会，基礎統計学I 統計学入門
 - 多くの先生がこれで学んだ標準的教科書
 - 手元にぜひ置いておきたい一冊
 - この講義で扱わない箇所も多い

- 参考書

- 逸見功，BLUE BACKS 統計ソフト「R」超入門
 - Rコマンドの解説書



- Massive Open Online Courses (MOOCs)
 - Coursera, edX など有名大学のコースを動画で受講
- gacco … JMOOC（日本版MOOC）のひとつ
 - 統計学II

https://lms.gacco.org/courses/course-v1:gacco+ga047+2020_10/about

- カーン・アカデミー (Khan Academy)
 - 日本語版はコンテンツ少ないので英語版お勧め
YouTube で khan academy statistics など検索

