



茨城大学農学部
College of Agriculture, Ibaraki University



農学分野のデータサイエンス 教育のあり方

茨城大学農学部・特任教授
統計数理研究所・名誉教授
(独)統計センター 特別研究員

田村義保

関東・首都圏ブロック農学分野
データサイエンス教育・AIの教育、研究の展開普及

2022年3月4日

オンライン





自己紹介



1975年 東京工業大学理学部物理学科卒業
 1980年 東京工業大学大学院理工学研究科物理学専攻博士課程修了(理学博士)
 1980年 日本学術振興会・奨励研究員
 1981年 統計数理研究所・研究員
 1985年 統計数理研究所・助手(共同利用研究所への組織変更に伴う職名変更)
 1986年 統計数理研究所助教授・総合研究大学院大学助教授(兼任1989年)
 1997年 統計数理研究所教授・総合研究大学院大学教授(兼任)
 1998年 統計数理研究所・技術課長(2000年まで)
 2000年 統計数理研究所・統計計算開発センター長(2004年まで)
 2004年 統計数理研究所・副所長(2010年まで)
 2010年 統計数理研究所・データ科学研究系主幹(2011年まで)
 2011年 統計数理研究所・副所長(2017年まで)
 2017年 統計数理研究所・データ科学研究系主幹(2018年まで)
 2018年 統計数理研究所・総合研究大学院大学 定年退職 名誉教授
 2018年 統計数理研究所・特任教授(2021年まで)
 2018年 独立行政法人統計センター 統計技術研究課・特別研究員
 2021年 茨城大学農学部・特任教授



<https://www.hpcwire.jp/archives/20001>



希望する数学知識

最新の課程の数I、II、A、B

分かっていて欲しい数学知識

微分が最大(小)値を求めるための演算であること。
積分が面積を求める演算であること。

Taylor展開

確率密度関数の積分が確率になること。
データは確率変数の実現値であること。

大学で習う？



分かっているとうれしい数学知識

ベクトル・行列・行列式の知識
行列の積や転置行列の概念

逆行列の意味
固有値、固有ベクトルの意味

QR分解などの数値計算
最適化(少なくとも未定乗数法)

中学や高校で数学が好きではなくなった学生に、
数学の面白さ、重要性を教えて欲しい。少なくとも、
数式を含む教科書などを読む元気を与えて欲しい。

大学で習う？

というより教えて欲しい



計算機リテラシー

WindowsパソコンかMacが使える。
WindowsとMacで文字コードや改行コードが違うことを理解している。

Rでのデータの読み込み方を知っている。
簡単なRコマンドを知っている。

後はやっている内に覚えることができる。

RではなくPythonでも良い。
Excelが使える?

計算機は少々の操作ミスでは壊れない。クリックするだけのブラックボックス解析では、解析結果の分析はできないと思っている。結果の分析が正しくできるような教育で良いのですが。

Rの出力を理解するのは結構難しい



データサイエンス教育で一番重要なものはサンプルデータです。

データのアノテーションやメタデータと解析例は必ずつけて欲しいものです。

次ページからいくつかのサイトを紹介しています。初等中等教育のためのページも上げていますが、大学教育のためにも役立つと思います。

残念ながら、観測、観察、計測したような農学関係のデータは農水省の統計データ以外には見当たらなかったように思います。



日本統計学会
統計教育委員会
「資料の活用」・「データの分析」のための教材サイト
<https://estat.sci.kagoshima-u.ac.jp/cse/shido/>

日本統計学会
統計教育委員会
Special Committee of Statistical Education

「資料の活用」・「データの分析」のための教材サイト

このページでは、新しく始まる小学校・中学校・高等学校の新学期指導要領における統計に関する授業の参考になる情報を発信します。

参考サイト

- センサス@スクール (パンフレット 図)
- 科学の道 toolbox (理科ねっとわーく) (パンフレット 図)
- なるほど統計学園
- 統計関連学会連合 統計教育推進委員会
- 統計グラフ教育研究部会



JDSSP 高等学校データサイエンス教育研究会

<https://ds-education.com/category/data-science-text/page/2/>

<https://ds-education.com/data-science-text/174/>

The screenshot shows a web browser window displaying the JDSSP website. The page title is 'JDSSP高等学校データサイエンス教育研究会'. The main content area is titled 'データ収集一覧' (Data Collection Overview) and lists various data sources and resources. The right sidebar contains a '最近の投稿' (Recent Posts) section and a 'カテゴリー' (Categories) section.

データ収集一覧
2020年10月23日

- DATA GO.JP (日本政府全体のデータ一覧サイト)
- e-Stat (政府統計の総合窓口)
- 統計GIS (地図で見る統計)
- RESAS (リーサス) 地域経済分析システム (データがすぐに可視化できるサイト)
- SSDSE (教育用標準データセット) 統計データ分析
- miripo (ミクロデータ利用ポータルサイト)
- 統計ダッシュボード (総務省統計局)
- なるほど統計学園
- 国土交通省 国土地理院
- 国、都道府県、政令指定都市「統計年鑑(統計書、県勢要覧、統計年報)」まとめさいと
- 各府省及び独立行政法人等のページ (国の省庁ごとのデータリンクサイト)

最近の投稿

- 理数系教員統計・データサイエンス授業力向上研修集会(愛媛)
- 小学5・6年向け統計教材
- 高等学校教員向け統計研修会(神戸大学数理データサイエンスセンター)
- 【12月17日】第13回 JDSSPデータサイエンス教育研究会
- 【11月17日】第12回 JDSSPデータサイエンス教育研究会

カテゴリー

- DS教材
- 研修・コンペ
- 研究会



統計教育推進委員会

<https://statds.k-junshin.ac.jp/statedu/>

The screenshot shows a web browser window displaying the homepage of the Statistical Education Promotion Committee. The page has a yellow header with the title '統計教育推進委員会' and a red sidebar menu on the left. The main content area is white with a red border. The sidebar menu includes links to 'INDEX', '活動案内', '資料', '委員会構成', '統計学分野の教育課程編成上の参照基準', and '教育教材サイト'. The main content area features a '活動案内' section with a paragraph about the committee's mission and a 'What's New' section with a list of recent events. The right sidebar contains a 'お問い合わせ' section with contact information and a '関連リンク' section with links to related projects.

統計教育推進委員会

INDEX

- 活動案内
- 資料
- 委員会構成
- 統計学分野の教育課程編成上の参照基準
- 教育教材サイト

活動案内

ご挨拶

統計関連学会連合では、広い意味での統計教育を推進するために統計教育推進委員会を発足させました。これは、統計科学の専門家集団である統計関連学会が小学校から社会人に至る各段階での統計教育に積極的に関わっていかうと考えたためです。各段階での統計教育に対する考え方、実際に教育を実践するために必要なデータ集などを掲載します。今後さらに中身を充実させていきます。皆さんのお考えなどを委員会メンバーにお伝え下さい。

委員長 渡辺 美智子（康徳義塾大学）

What's New

- 2016.8.17 文部科学省へ統計関連学会連合から「高度情報化・データサイエンス社会における国民の科学的判断力・探究力強化に向けた新科目『数理探究(仮称)』内容への要望書」を提出されました。（[要望書](#)）
- 2012.9.3 文部科学省大学間連携共同教育推進事業に 8大学(東京大学,大阪大学,総合研究大学院大学,青山学院大学（代表校）,多摩大学,立教大学,早稲田大学,同志社大学）、統計関連学会連合の6学会、8団体(大学入試センター,日本アクチュアリー会,日本科学技術連盟,日本銀行,日本経済団体連合会,日本製薬工業協会,日本統計協会,日本マーケティング・リサーチ協会) が連携して申請した「データに基づく課題解決型人材育成に資する統計教育質保証」事業が9月3日に選定されました。
- 2013.3.25 統計科学のための電子図書システムが公開されました。
- 2013.9.18 日本学術会議より数理科学分野の「大学教育の分野別質保証のための教育課程編成上の参照基準」が公表されました。この参照基準では、「数理科学」を数学、統計学、応用数学の3分野からなるものと定義しています。統計学についても、かなり詳しく書かれています。

お問い合わせ

本会に対する問い合わせは、事務局 までお願いいたします。
statedu@jfssa.jp

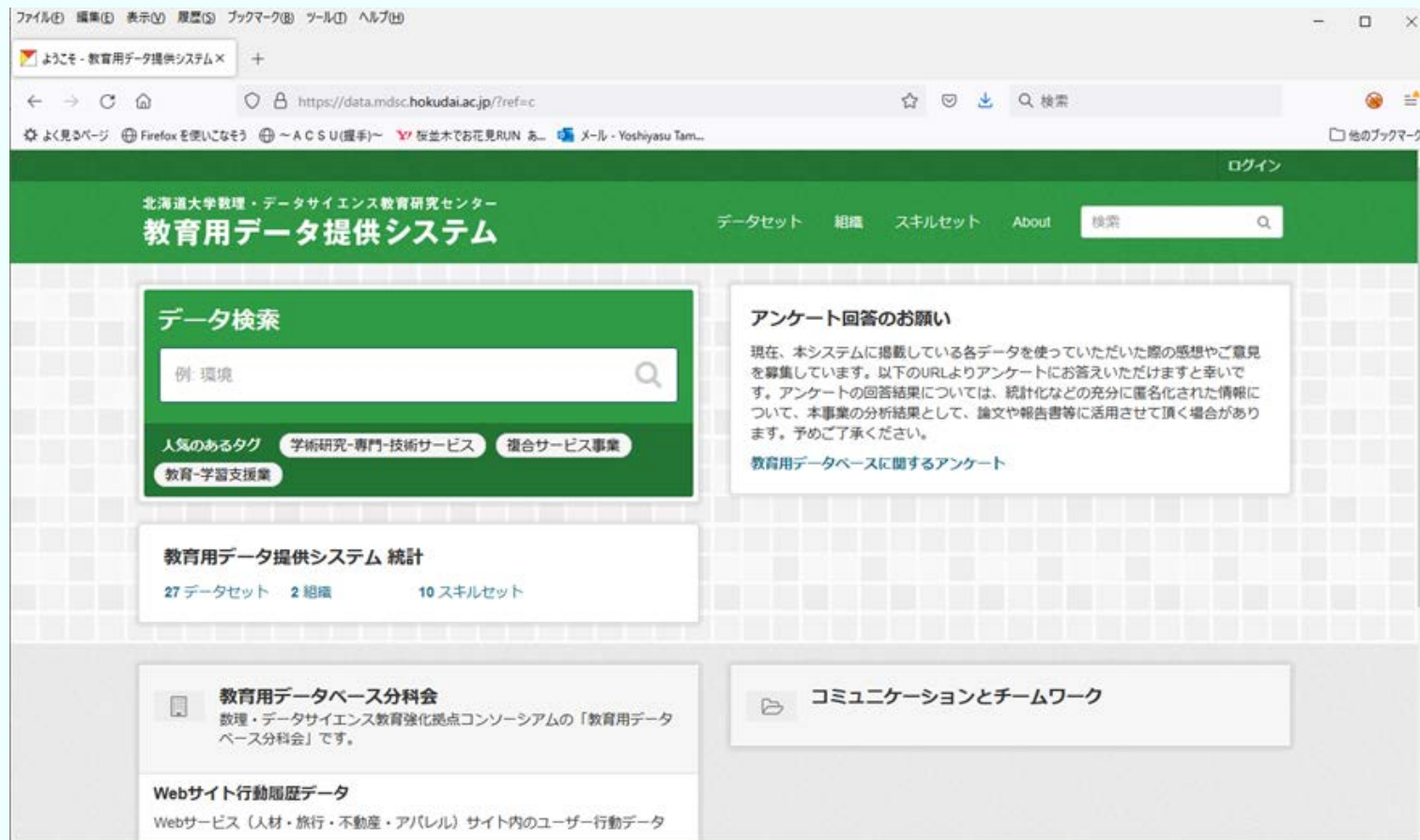
関連リンク

- センサス@スクール（日本版）
- International CensusAtSchool Project [New Zealand]
- International CensusAtSchool Project [United States]
- 総務省統計局統計学ウェブサイト



北海道大学数理・データサイエンス教育研究センター 教育用データ提供システム

<https://data.mdsc.hokudai.ac.jp/?ref=c>



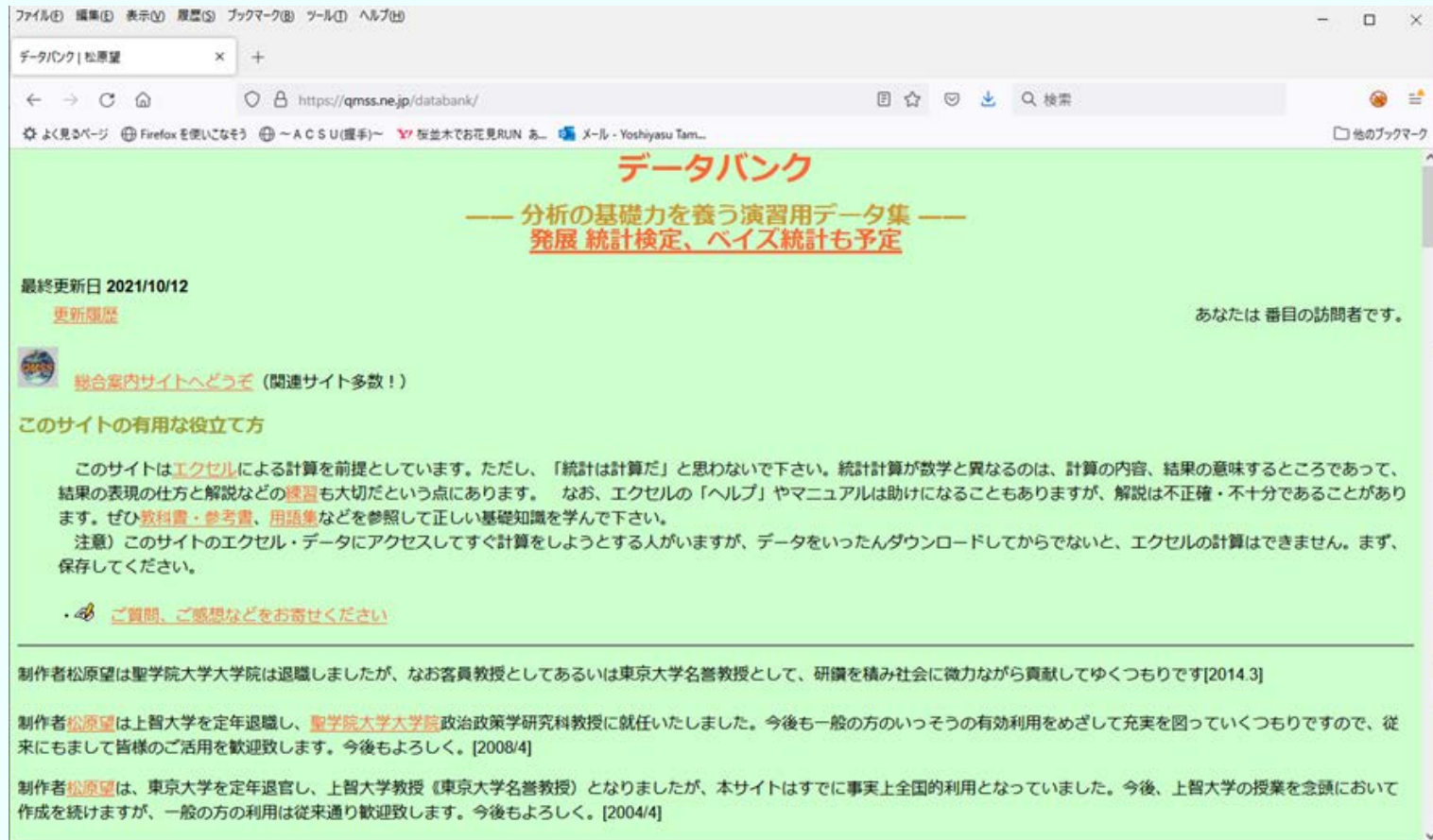
The screenshot shows a web browser window displaying the "教育用データ提供システム" (Educational Data Provision System) website. The page has a green header with the title and navigation links: "データセット" (Data Set), "組織" (Organization), "スキルセット" (Skill Set), and "About". A search bar is located on the right side of the header. The main content area is divided into several sections:

- データ検索** (Data Search): A search bar with the example text "例: 環境" (Example: Environment) and a magnifying glass icon. Below it are tags for "人気のあるタグ" (Popular Tags), including "学研究-専門-技術サービス" (Research-Specialty-Technical Service), "複合サービス事業" (Composite Service Business), and "教育-学習支援策" (Education-Learning Support Strategy).
- アンケート回答のお願い** (Request for Survey Responses): A text block explaining that the system collects user feedback and requests responses to a survey. It includes a link to "教育用データベースに関するアンケート" (Survey regarding the Educational Database).
- 教育用データ提供システム 統計** (Educational Data Provision System Statistics): A section showing statistics: "27 データセット" (27 Data Sets), "2 組織" (2 Organizations), and "10 スキルセット" (10 Skill Sets).
- 教育用データベース分科会** (Educational Database Subcommittee): A section mentioning the "数理・データサイエンス教育強化拠点コンソーシアムの「教育用データベース分科会」" (Mathematics and Data Science Education Enhancement Hub Consortium's "Educational Database Subcommittee").
- Webサイト行動履歴データ** (Website Activity History Data): A section mentioning "Webサービス (人材・旅行・不動産・アパレル) サイト内のユーザー行動データ" (Web Services (Human Resources, Travel, Real Estate, Apparel) User Activity Data within the Site).
- コミュニケーションとチームワーク** (Communication and Teamwork): A section with a folder icon.



松原望先生 データバンク

<https://qmss.ne.jp/databank/>

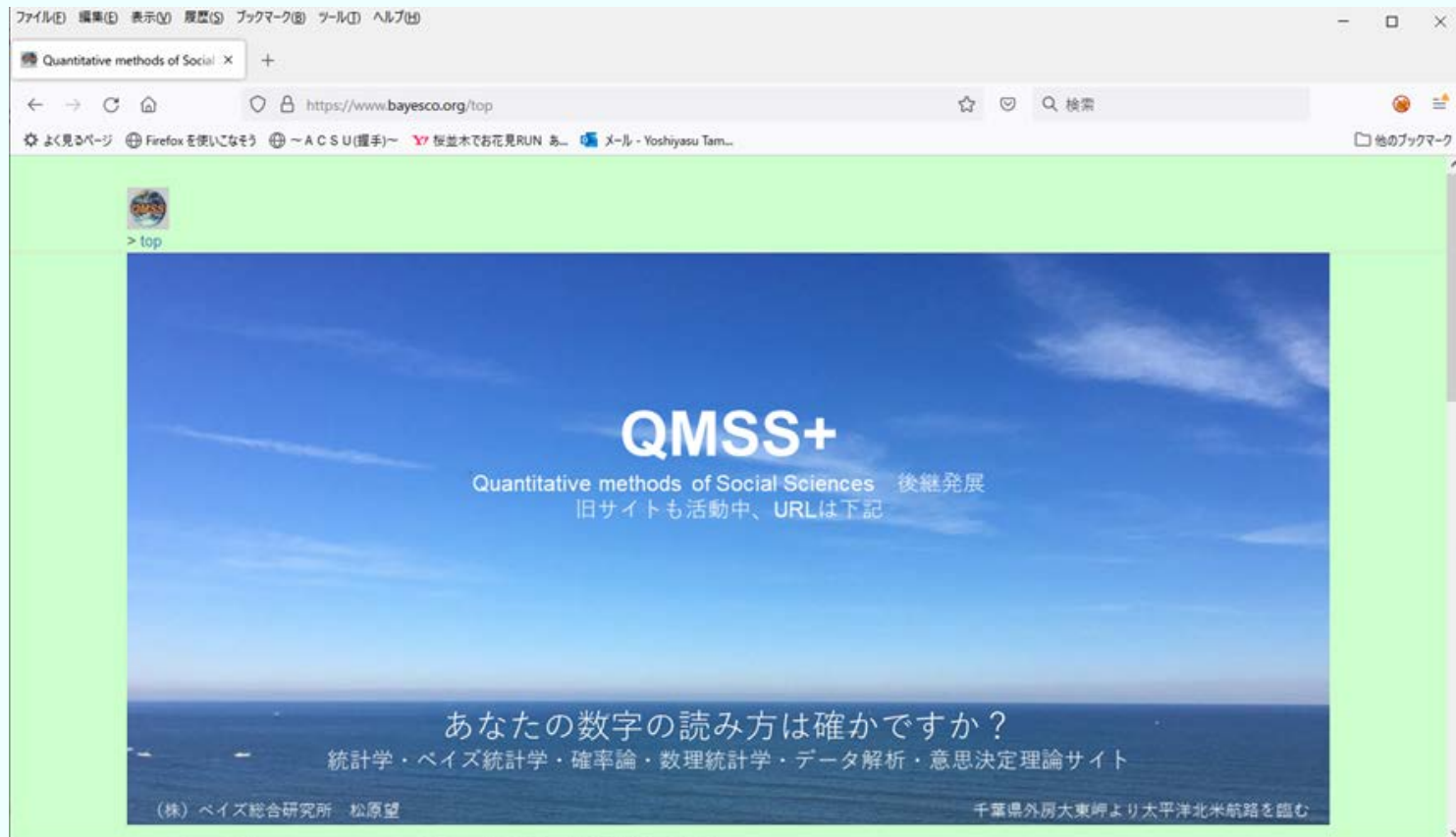


松原先生 総合案内サイト <https://qmss.ne.jp/portal/>



松原先生 新サイト

<https://www.bayesco.org/top>



わかりやすい統計学 データサイエンス基礎 松原望・森本栄一 丸善出版 2021年11月25日



統計データ分析コンペティション

<https://www.nstac.go.jp/statcompe/>

統計データ分析コンペティション

SSDSE

- 統計データ分析コンペティション
 - 論文募集要項
 - 受賞論文
 - 過去の受賞論文・募集要項
- SSDSE (教育用標準データセット)
 - 過去のSSDSE
 - SSDSEの利活用事例

トップページ (統計データ分析コンペティション)

統計データ分析コンペティション

高校生、大学生等を対象に、地域別の統計をまとめた SSDSE (教育用標準データセット) を用いた統計データ分析の論文を募集し、そのアイデアと解析力を競うコンペティションです。

平成30年度から毎年開催しており、令和3年度(2021年度)に第4回の「統計データ分析コンペティション2021」を開催しました。

2022年度(第5回)の募集要項等は、2022年3月末頃に掲載の予定です。

(注) SSDSEは、データサイエンス教育のための汎用素材として、作成・公開しているデータセットです。

2022年1月1日 2021年度の受賞論文の、月刊誌『統計』への連載が始まりました(1月号～8月号、日本統計協会発行)。

2021年12月8日 統計データ分析コンペティションの受賞論文全文と審査会総評を掲載しました。

2021年10月18日 総務省の報道発表「第6回 地方公共団体における統計データ利活用表彰」の受賞団体及び「統計データ分析コンペティション2021」の受賞者の決定(PDF: 160KB)により、コンペティションの受賞者及び受賞論文が発表されました。

SSDSE が標準データセットです



政府統計総合窓口

<https://www.e-stat.go.jp/>



建設工事受注動態統計調査の結果もあります。あまり知られていないかもしれませんが、基幹統計のひとつの令和2年度の国民生活基礎調査は中止になっています。



考えているコース

農学のための統計入門 記述統計学
グラフ・表による情報集約

1. ヒストグラム
2. 累積相対度数グラフ(累積百分率グラフ)
3. 経験分布グラフ
4. ボックスプロット
5. 円グラフ
6. 帯グラフ
7. 折れ線グラフ
8. チャーノフの顔グラフ
9. 時系列プロット
10. QQ プロット
11. ローレンツ曲線
12. 散布図

1. 度数分布表
2. 相関表
3. 分割表

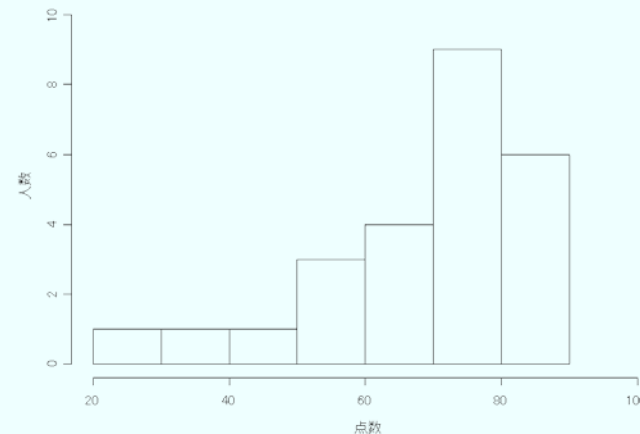


表や図を最初を書くべきである。

入試	国語	社会	数学	理科	英語
78	73	70	74	79	71
65	29	33	55	79	84
79	71	68	72	64	97
86	74	91	79	76	100
67	52	56	58	60	85
92	77	92	96	88	98
78	60	85	66	66	88
88	81	91	73	63	95
78	61	84	72	78	92
90	70	75	81	67	96
74	53	70	73	51	92
82	69	64	96	57	97
96	87	89	90	85	100
91	83	75	96	81	98
73	76	61	67	57	86
89	87	82	78	85	97
85	77	80	78	70	94
59	38	43	45	12	96
80	67	73	78	67	95
91	83	77	80	67	100
62	47	61	56	21	95
80	70	62	88	51	96
80	81	51	63	66	92
57	51	16	36	48	84
74	76	71	77	75	79

階級	階級値	度数	相対度数 (%)	累積度数	累積相対 度数(%)
20-30	25	1	4	1	4
30-40	35	1	4	2	8
40-50	45	1	4	3	12
50-60	55	3	12	6	24
60-70	65	4	16	10	40
70-80	75	9	36	19	76
80-90	85	6	24	25	100
計	-	25	100	-	-

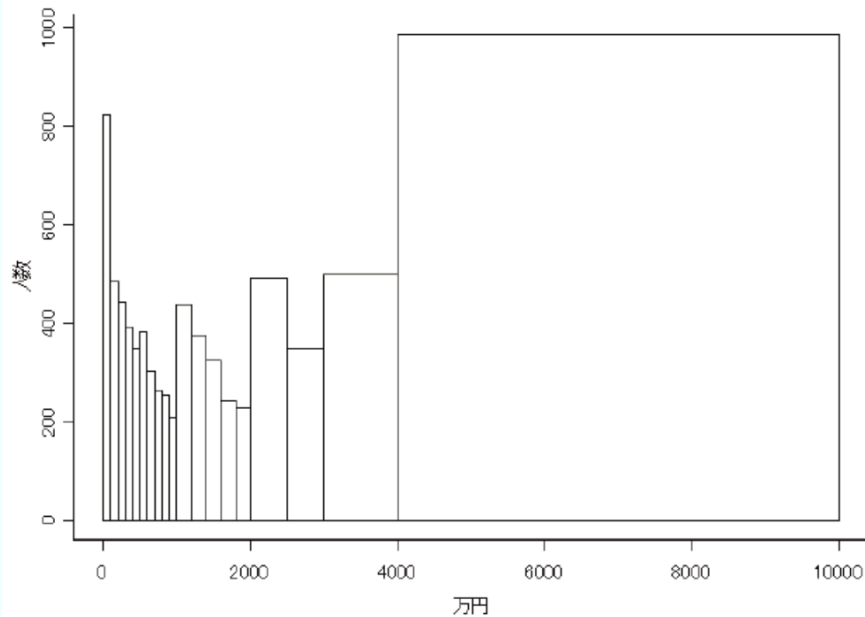
国語の度数分布



上のようなデータをCase & Variable型という。相対度数や累積相対度数は確率論と密接に関係している。このあたりを理解できれば、深い解析が可能になる。

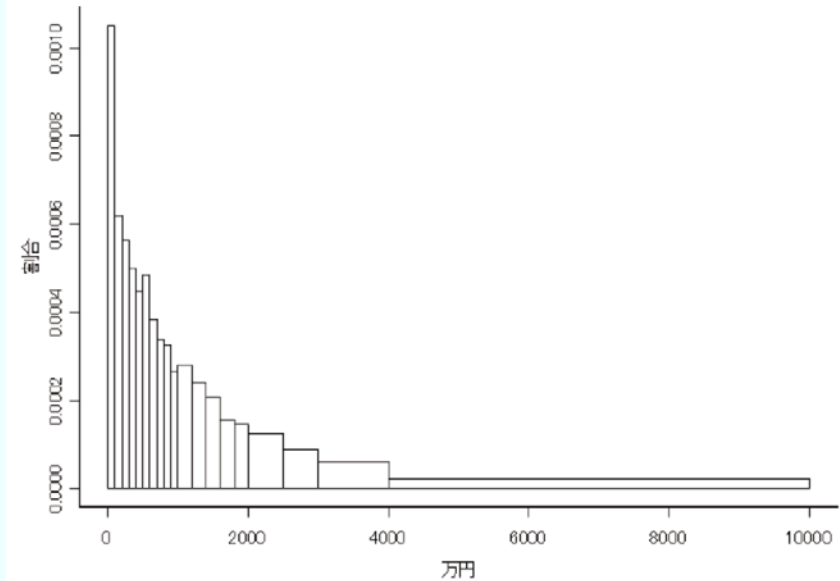


貯蓄現在高(不適切ヒストグラム)



高さを度数にしています。階級幅が一定ならばこれでもよいですが。。

貯蓄現在高(適切ヒストグラム)



面積が度数に比例するように描いています。棒グラフの面積の合計は1です。

ヒストグラムについてふれているのは、ヒストグラムが確率密度関数の近似であるため。また、データの分布を表しているため。ヒストグラムは小学校から習っているので、かえって、変な「すりこみ」があるように思う。



```
hist(iba2002.dat[1:7852],breaks=c(0,seq(99.9,999.9,100),  
seq(1199.9,1999.9,200),  
2499.9,2999.9,3999.9,10000),freq=T,  
xlab="万円",ylab="人数",main="貯蓄現在高(不適切ヒストグラム)",col=0,  
ylim=c(0,1000))
```

```
dev.copy2eps(file="fig-hist01.eps",family="Japan1")
```

```
hist(iba2002.dat[1:7852],breaks=c(0,seq(99.9,999.9,100),  
seq(1199.9,1999.9,200),  
2499.9,2999.9,3999.9,10000),xlim=c(0,10000),  
xlab="万円",ylab="割合",main="貯蓄現在高(適切ヒストグラム)",col=0,freq=F)
```

```
dev.copy2eps(file="fig-hist02.eps",family="Japan1")
```



考えているコース

農学のための統計入門 記述統計学
数値による情報集約

1. 総計値
2. 平均値
3. 最大値・最小値
4. パーセンタイル
5. メディアン(中央値)
6. 第1四分位数
第3四分位数
7. モード
8. 範囲
9. 四分位範囲
10. 分散・標準偏差

11. 平均偏差
12. 四分位偏差
13. 変動係数
14. 四分位変動係数
15. 歪度
16. 尖度
17. 共分散、相関



- 名義尺度(nominal scale)
 - ✓ いくつかのものを区別するために、一連の数値を与える
 - ✓ ラベルとしての意味しかない
 - ✓ 例 性別 企業業種
- 順序尺度(ordinal scale)
 - ✓ 順序を表すデータ
 - ✓ 例 (良い、普通、悪い)
- 間隔尺度(interval scale)
 - ✓ 計量データの中で、与えられた数値が、差の等価性を持つが
 - ✓ 絶対的な原点を持たない場合
 - ✓ 例 温度(セ氏)、試験の点数
- 比率尺度(ratio scale)
 - ✓ 間隔尺度の性質に加え、原点が一意に決まっている場合
 - ✓ 例 長さ、重さ、時間等の物理量

- 質的データ
 - ✓ カテゴリカルデータ : 名義尺度
 - ✓ 順序データ : 順序尺度
- 量的データ
 - ✓ 計量データ : 間隔尺度 比率尺度
 - ✓ 計数データ

どのような尺度を持つかによって、使える統計手法が異なる。尺度のことを書いている日本語教科書は少ないと思う。講義評価は順序尺度なので平均点を出すのは本当は不適切である。



標本平均 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Rコマンド(1)

```
> setwd("d:/Rwd")  注意:作業ディレクトリを指定。DドライブのRwdという  
                    フォルダを想定  
> baby <- matrix(scan("baby.dat"),ncol=5,byrow=T)  d:¥Rwd¥baby.datを  
                                                    読み込み  
> baby  
> mean(baby[,1])
```

大銀行に所属する方から、『平均とは足して2で割るものだと思っている人が多い』という
ことを聞いて驚いたことがある。標本平均より、メディアンやモードの方が代表値としてよ
いということを知っていた方がよい。いきなり、平均を計算してはいけない。シグマ記号が
あるだけで統計的データ解析を忌避する人がある。



> baby の結果

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	3087	48	28	304	0
[2,]	3229	52	24	286	1
[3,]	3204	61	33	273	1
[4,]	3346	58	30	295	0
[5,]	3579	56	21	290	0
[6,]	2325	46	26	262	0
[7,]	3159	55	30	318	1
[8,]	3589	63	37	298	0
[9,]	2969	52	25	299	1
[10,]	2819	40	22	313	0
[11,]	3191	59	34	285	1
[12,]	3346	57	28	306	0
[13,]	2444	45	30	291	1
[14,]	3662	64	21	274	0
[15,]	3241	53	29	283	0

標本平均の計算

```
> mean(baby[,1])  
[1] 3146  
> apply(baby,2,mean)  
[1] 3146.00000 53.93333  
27.86667 291.80000 0.40000
```



標本分散 $v^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2$ 二乗の平均ー平均の二乗の式は使ってはいけない

不偏分散 $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

不偏分散を用いるのは、その期待値が母分散(真の値)に等しいため
しかし、分散の最尤推定値は標本分散である。AIC(赤池情報量規準)で用いるのは最尤推定値であることに注意

Rコマンド(2)

```
> var(baby[,1])
```

不偏分散の計算

```
> var(baby[,1])
```

```
[1] 147146.4
```

標本分散の計算

```
> (length(baby[,1])-1)*var(baby[,1])/length(baby[,1])
```

```
[1] 137336.7
```

$n - 1$ で割る方が統計的によいわけではない。不偏性という観点からよいだけ。



3. 最大値・最小値

$$\text{最大値} = x_{(n)}$$

$$\text{最小値} = x_{(1)}$$

パーセンタイルが数値による情報集約(縮約)では一番重要であるように考える。

4. パーセンタイル

データを小さい順に並びかえたとき、小さい方から $100p\%$ ($0 < p < 1$) のところにある値を $100p$ パーセンタイルという。

度数分布表の一部

7-8	7.5	103	10.3	143	14.3
8-9	8.5	186	18.6	329	32.9
9-10	9.5	216	21.6	545	54.5
10-11	10.5	163	16.3	708	70.8
11-12	11.5	135	13.5	843	84.3

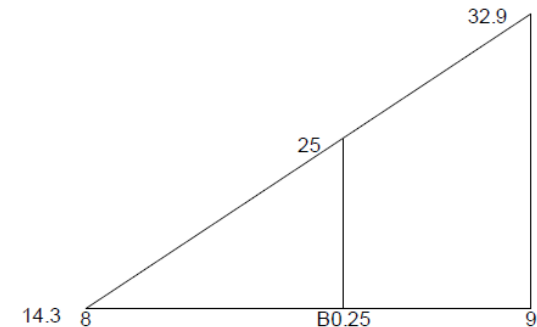


度数分布表から求めた 25 パーセンタイル、50 パーセンタイル、75 パーセンタイル

$$\frac{B_{0.25} - 8}{9 - 8} = \frac{25 - 14.3}{32.9 - 14.3} \quad B_{0.25} = 8.575269$$

$$\frac{B_{0.5} - 9}{10 - 9} = \frac{50 - 32.9}{54.5 - 32.9} \quad B_{0.5} = 9.791667$$

$$\frac{B_{0.75} - 11}{12 - 11} = \frac{75 - 70.8}{84.3 - 70.8} \quad B_{0.75} = 11.31111$$



全データから求めた 25 パーセンタイル、50 パーセンタイル、75 パーセンタイル

0%	25%	50%	75%	100%
5.2	8.7	9.8	11.3	20.3

補間法に過ぎませんが、講義では三角形の相似形を使って説明しています。



今は小学校の算数
で第1四分位数の求
め方が説明されて
いますが、教科書に
よって違いがあるそ
うです。

パーセンタイル (0%から 100%、5%きざみ)

5.2	7.2	7.7	8.1	8.4
8.7	8.9	9.2	9.3	9.6
9.8	10.1	10.34	10.7	11
11.3	11.6	12.1	12.7	13.705
20.3				

補間法に過ぎ
ない。

パーセンタイル(確率点)を求める 関数

```
> quantile
function(x, probs = seq(0, 1, 0.25))
{
  if(any(probs < 0 | probs > 1))
    stop("Probabilities must be between 0 and 1 inclusive")
  if(any(is.na(x)))
    return(NA)
  names(x) <- NULL
  n <- length(x)
  order <- 1 + (n - 1) * probs
  low <- pmax(floor(order), 1)
  high <- pmin(low + 1, n)
  x <- sort(x, unique(sort(c(low, high))))
  order <- order %% 1
  (1 - order) * x[low] + order * x[high]
}
```




モードの計算法

モード(最頻値)は、普通は、単峰のヒストグラムで、一番高い階級の階級値となります。ここで、説明している求め方は、一番高いところの両隣の高さを考慮した求め方です。

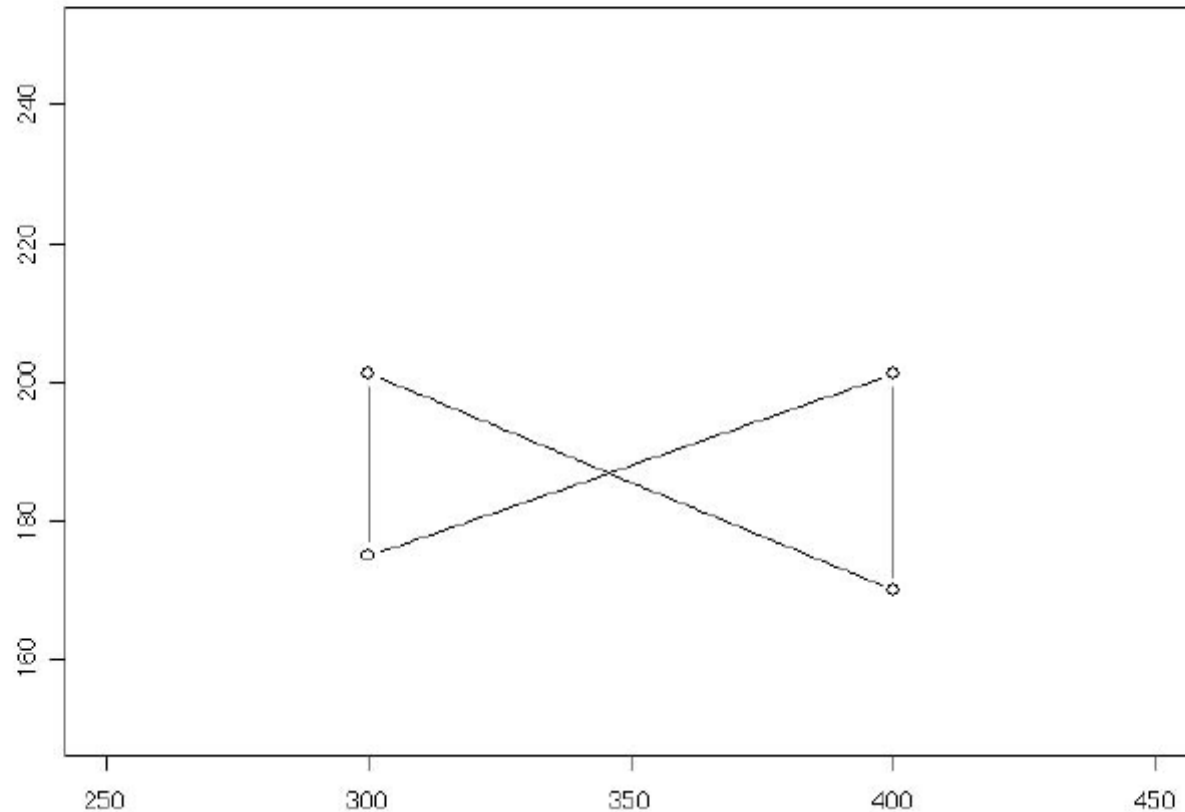
金融広報中央委員会「家計の金融資産に関するアンケート」平成15年度
金融資産の保有状況

階級			階級値	度数
0	-	100	50	150
100	-	200	150	193
200	-	300	250	175
300	-	400	350	201
400	-	500	450	170
500	-	600	550	168
600	-	700	650	129
700	-	800	700	116
800	-	900	850	121
900	-	1000	950	96



モードの計算

モードの計算の仕方





ヒストグラムからモードは300以上400未満の値となる。モードを $300 + x$ とすると、三角形の相似形の性質を使って x は次のように計算できる。

$$\frac{x}{100 - x} = \frac{201 - 175}{201 - 170}$$

$$31x = 26(100 - x)$$

$$57x = 2600$$

$$x = 45.61404$$

したがって、モードは345.6万円である。



考えているコース

農学のための統計入門 推測統計学
推測統計学 区間推定 仮説検定 等

1. 確率変数の和の分布
2. 平均値の区間推定
3. 分散の区間推定
4. 比率の区間推定

5. 平均値の検定
6. 分散の検定
7. 平均値の差の検定
8. 分散の比の検定
9. 比率の差の検定

10. 独立性の検定
11. 分散分析
12. 単回帰



連続確率変数と確率密度関数

連続確率変数 X : 変数 X がある特定の値をとることが偶然に支配されており（我々が定めることができません）かつ、ある特定の範囲 $(x, x + \Delta x]$ の値をとる確率が与えられている時、その変数を連続確率変数と呼ぶ。

確率密度関数 $f(x)$: 連続確率変数 X がある特定の範囲 $(x, x + \Delta x]$ の値をとる確率が

$$P(x < X \leq x + \Delta x) = f(x)\Delta x$$

で表される場合、 $f(x)$ を確率変数 X の確率密度関数と言う。次式の方が数学的には厳密です。

$$\lim_{\Delta x \rightarrow 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} = f(x)$$



標本平均

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

不偏分散

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

の分布

母集団が平均 μ 、分散 σ^2 の正規分布に従う場合、 n 個のデータ x_i , $i = 1, \dots, n$ から計算される標本平均 $\bar{x}$ は平均 μ 、分散 $\frac{\sigma^2}{n}$ の正規分布に従う。(標本自体の分散をデータ数で割ったものが標本平均の分散となるということです。)

母集団が平均 μ 、分散 σ^2 の正規分布に従う場合、 n 個のデータ x_i , $i = 1, \dots, n$ から計算される不偏分散を s^2 とする時、

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

は自由度 $n - 1$ のカイ二乗分布 (χ 二乗分布) に従う。

実験データは普通は一組しかありません。何回も同じ実験を繰り返したら、標本平均の分布が正規分布になるということです。仮想的に実験を繰り返したのと同じようにする統計的手法として、ブートストラップ法(Efron)があります。



母平均 μ の信頼区間 $100(1 - 2\alpha)\%$ の区間推定(母分散が既知)

$$\bar{x} - z_{\alpha} \times \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + z_{\alpha} \times \sqrt{\frac{\sigma^2}{n}}$$

$\alpha = 0.05$ の場合は $z_{0.05} = 1.64$ 、 $\alpha = 0.025$ の場合は $z_{0.025} = 1.96$ であることから、信頼区間90%および95%の区間推定

$$\bar{x} - 1.64 \times \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.64 \times \sqrt{\frac{\sigma^2}{n}}$$

$$\bar{x} - 1.96 \times \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}$$

信頼区間90%というのは同じ推定を100回繰り返すことができたなら、その内、90回は、不等式の範囲に母平均(真の値)が入るという意味を持ちます。



母分散 σ^2 の信頼区間 $100(1 - 2\alpha)\%$ の区間推定

$$\chi_{1-\alpha}^2(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{\alpha}^2(n-1)$$

であることを用いて次のようになります。

$$\frac{(n-1)s^2}{\chi_{\alpha}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha}^2(n-1)}$$

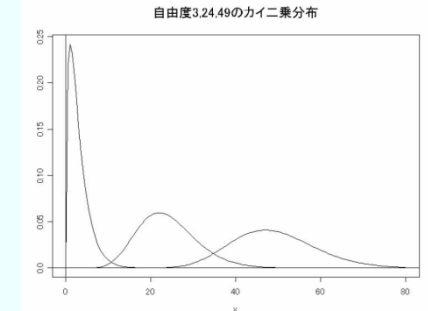
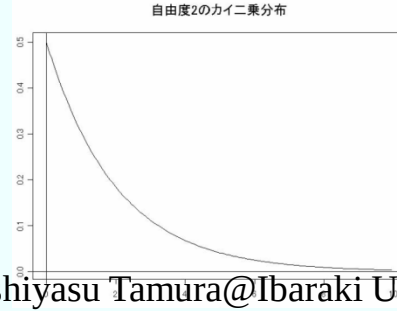
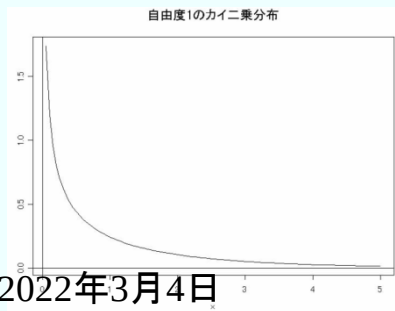
自由度 k のカイ二乗分布の
確率密度関数

$$f(x) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad (x \geq 0), \quad 0 \quad (x < 0)$$

自由度 k のカイ二乗分布の
期待値と分散

$$E[X] = k$$

$$V[X] = E[(x - k)^2] = 2k$$



2022年3月4日

Yoshiyasu Tamura@Ibaraki Univ.



平均値の差の検定

2つの標本の平均値の差の検定を行うための検定統計量について説明しています。

平均値の差の検定

母集団 A が母平均 μ_1 、母分散 σ_1^2 の正規分布に従うとする。

$$X_i \sim N(\mu_1, \sigma_1^2), i = 1, 2, \dots, n$$

標本平均に対応する確率変数。

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim N(\mu_1, \frac{\sigma_1^2}{n})$$

母集団 B が母平均 μ_2 、母分散 σ_2^2 の正規分布に従うとする。

$$Y_i \sim N(\mu_2, \sigma_2^2), i = 1, 2, \dots, m$$

標本平均に対応する確率変数。

$$\bar{Y} = \frac{\sum_{i=1}^m Y_i}{m} \sim N(\mu_2, \frac{\sigma_2^2}{m})$$

標本平均の差に対応する確率変数

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$



平均値の差の検定

平均値の差の検定 続き

分散が既知の場合 標準化

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$

不偏分散に対応する確率変数。

$$S_1^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

$$S_2^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m - 1}$$

プーリングした分散

$$S^2 = \frac{(n - 1)S_1^2 + (m - 1)S_2^2}{n + m - 2}$$

t 値

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2}{n} + \frac{S^2}{m}}} \sim t - \text{distri}(n + m - 2)$$



平均値の差の検定

帰無仮説 $\mu_1 = \mu_2$ 母分散既知の場合は次の検定統計量が標準正規分布に従うことを用いて行うことができる。

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$

帰無仮説 $\mu_1 = \mu_2$ 母分散未知だが等しいとできる場合は次の検定統計量が自由度 $n + m - 2$ の t 分布に従うことを用いて行うことができる。

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s^2}{n} + \frac{s^2}{m}}} \sim t - \text{distri}(n + m - 2)$$



平均値の差の検定

分散が未知であるが等しいとできる場合についてのみ説明します。帰無仮説は母平均が等しいです。対立仮説は標本Aの母平均の方が大きい、小さいと標本Aと標本Bの母平均は異なるの3種類考えられますが、等しくないという対立仮説についてのみ説明します。

3. 対立仮説： $\mu_1 \neq \mu_2$

検定統計量の値を t とする時)、自由度 $n + m - 2$ の t 分布で $|t|$ より大きい $-|t|$ より小さな値をとる確率を p 値という。

$$P(T > |t|) + P(T < -|t|) = 2 \times P(T > |t|) = p$$

R では

$$2 * (1 - pt(|t|, n + m - 2))$$

で計算することができる。Excel では

$$2 * (1 - T.DIST(|t|, n + m - 2, 1))$$

$$2 * T.DIST.RT(|t|, n + m - 2)$$

で計算することができる。



平均値の差の検定

Rでは二つの標本の分散が等しいとできる場合は、`t.test( , var.equal=T)`を用いて行うことができる。`t.test`のデフォルトは両側検定で、二標本の分散が等しく無いである。

```
t.test —  
t.test(x, y = NULL,  
       alternative = c("two.sided", "less", "greater"),  
       mu = 0, paired = FALSE, var.equal = FALSE,  
       conf.level = 0.95, ...)
```



平均値の差の検定

ACとACSの播種後31日目の草丈については次のようになりました。

播種後31日の草丈 ACとACSの平均値のt検定（等分散）

```
> xx1 <- potato31h.df$ac  
> xx2 <- potato31h.df$acs  
> yy1 <- xx1[!is.na(xx1)]  
> yy2 <- xx2[!is.na(xx2)]  
> t.test(yy1, yy2, var.equal=T)
```

Two Sample t-test

```
data: yy1 and yy2  
t = -1.2097, df = 29, p-value = 0.2362  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -3.1391721  0.8058388  
sample estimates:  
mean of x mean of y  
 10.83333  12.00000
```




平均値の差の検定

ACとACSの播種後31日目の草丈についてExcelを用いると次のようになり、Rを用いた結果と一致します。

平均値の差の検定

AC	14	13	10	11	13
	10	11	12.5	10	11
	11	1	12	11	12

ACS	13.5	13.5	16	14	6.5
	13	12	10	14	8.5
	13	11	13	13	11
	10				

AC	ACS
標本平均の計算	標本平均の計算
標本平均: 10.83333	標本平均: 12
不偏分散: 8.845238	不偏分散: 5.666667
データ長: 15	データ長: 16
プーリング後不偏分散	7.201149

検定統計量の計算
t= -1.20968

p値 0.236171

平均値の差の検定

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

検定統計量

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

有意水準 5 %
α = 0.05
α/2 = 0.025
p = 2 * α/2 0.05

自由度n+m-2のt分布の上側 2.5 %点
2.04523

エクセルの計算式

$$t_{\alpha/2}(n+m-2) \leftarrow T.INV.2T(p, n+m-2), p = \alpha$$

判断基準
|t| > t_{α/2}(n+m-2): H₀を棄却
|t| ≤ t_{α/2}(n+m-2): H₀を容認

$$p = 2 * T.DIST.RT(|t|, n+m-2)$$

不偏分散

$$s_1^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$s_2^2 = \frac{\sum_{i=1}^m (y_i - \bar{y})^2}{m-1}$$

プーリングした分散

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$



考えているコース

農学のための統計入門 多変量解析
多変量解析等

1. 重回帰分析
2. 主成分分析
3. クラスタ分析
4. 判別分析
5. 因子分析
6. 時系列解析

7. 傾向スコア
8. 因果分析
9. 分類のための解析方法
10. カルバック・ライブラーダイバージェンス
11. Fisher情報行列
12. 赤池情報量規準



単回帰について

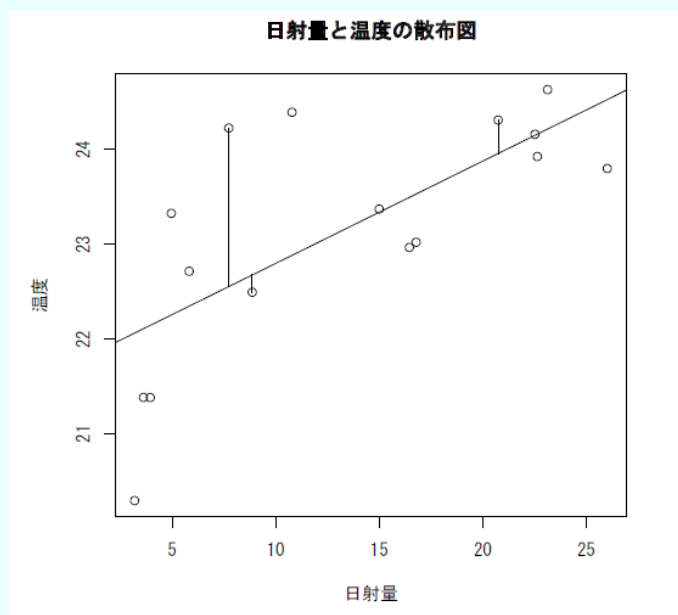
lm()を使うだけで単回帰分析は可能です。このデータでは気温(目的変数)の変化を日射量(説明変数)で説明・予測するために回帰分析を用いています。

```
2021年5月17日から6月1日 ————  
# 回帰分析  
> attach(summary01.df)  
> summary01.lm <- lm(temp ~ par)  
> summary(summary01.lm)  
Call:  
lm(formula = temp ~ par)  
  
Residuals:  
      Min       1Q   Median       3Q      Max   
-1.76874 -0.57804 -0.08556  0.38012  1.67416  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)      
(Intercept)  21.7177     0.4573   47.487  <2e-16 ***  
par           0.1080     0.0297    3.636   0.0027 **    
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.9283 on 14 degrees of freedom  
Multiple R-squared:  0.4857,    Adjusted R-squared:  0.449   
F-statistic: 13.22 on 1 and 14 DF,  p-value: 0.002696  
> detach(summary01.df)
```



単回帰について

3本しか描いていませんが、垂直な線が残差です。この二乗の和を最小にする傾きと切片を求めます。





単回帰について

残差二乗和を最小にする推定方法を最小二乗法と呼びます。 S を最小にする a と b を求めますが、数学的には、偏微分を用いて正規方程式を作ります。

単回帰分析 最小二乗法

p4 と p5 では日射量と気温の単回帰分析を行っています。R では `lm()` という関数で簡単にすることができます。単回帰分析は目的変数 y_i と説明変数 x_i があるとき、

$$y_i = a + bx_i + e_i, \quad i = 1, 2, \dots, n$$

の線形モデルを構築する解析方法です。ここで、 e_i は前ページの図では、点 (x_i, y_i) と $(x_i, a + bx_i)$ を結んでいる垂線を指します。 a と b の推定値を $\hat{a}$ 、 $\hat{b}$ とするとき、 x_i を用いた推定値 $\hat{y}_i = \hat{a} + \hat{b}x_i$ を得ることができます。 $e_i, i = 1, 2, \dots, n$ は互いに独立であることが仮定されています。この仮定は良く忘れられ、次に説明する最小二乗法で $\hat{a}$ と $\hat{b}$ を求めるだけで解析を終了すること多いですが、残差の分析、残差診断も行うことを勧めます。

$\hat{a}$ と $\hat{b}$ は次の残差二乗和 S を最小にするように決めます。

$$S = \sum_{i=1}^n (y_i - a - bx_i)^2$$



単回帰について

正規方程式は単に連立方程式です。文字式で解いていますが、推定値 $\hat{b}$ は共分散を説明変数の分散で割るという非常に美しい式になっています。推定値 $\hat{a}$ は元の直線の式から残差の項を除いただけです。

a 、 b で偏微分した値が0となるように決めます。

$$\sum_{i=1}^n 1 \cdot a + \sum_{i=1}^n x_i \cdot b = \sum_{i=1}^n y_i$$

$$\sum_{i=1}^n x_i \cdot a + \sum_{i=1}^n x_i^2 \cdot b = \sum_{i=1}^n x_i y_i$$

このような連立方程式を正規方程式といいます。簡単に解くことができ、次のようになります。

$$\hat{b} = \frac{v_{xy}}{v_x^2} = \frac{v_y}{v_x} r_{xy}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

ただし、説明変数、目的変数の分散を v_x^2 、 v_y^2 、説明変数と目的変数との共分散を v_{xy} としています。

$$v_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

$$v_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

$$v_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$



単回帰について

このデータの場合の計算は次の通りです。lm()を用いずに、理論式を用いています。

$$\hat{b} = \frac{v_{xy}}{v_x^2} = \frac{7.033508}{65.117419} = 0.1080127$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 23.15072 - 0.1080127 \times 13.26742 = 21.71767$$

となり、lm() で求めた結果と一致します。



単回帰について

分散分析をしています。

単回帰分析 分散分析

次に、回帰分析を行うことに意味があるかどうかを判断するために用いる分散分析について説明します。分散分析は複数の条件で実験を繰り返したような場合に、実験条件によって、結果が異なるかどうかを判断するために用いることが良く知られています。基本となる恒等式を示します。変動の分解と言われる式で、数値統計学ではいろいろな解析法の基本となっています。

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

総変動 SS_T がモデルによる変動 SS_A と誤差（残差）変動 SS_E に分解されるということである。

$$\begin{aligned} SS_T &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ SS_A &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ SS_E &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \end{aligned}$$



単回帰について

分散分析をしています。

本報告では単回帰を扱っているので、 $\hat{y}_i = \hat{a} + \hat{b}x_i$ であるが、説明変数が k 個の重回帰の場合は、

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1x_{1i} + \hat{b}_2x_{2i} + \cdots + \hat{b}_kx_{ki}$$

とするだけで、同じ関係が成り立つ。このため、分散分析の説明では、説明変数の個数を k と書くことにする。

残差 e_i が平均 0 分散 σ^2 の正規分布に従うものとします。

$$e_i \sim N(0, \sigma^2), i = 1, 2, \dots, n, iid$$

この時、 S_E^2 は σ^2 の不偏推定量になります。また、回帰分析することに意味がないという帰無仮説が真の場合には、次の S_A^2 も σ^2 の不偏推定量になります。

$$S_A^2 = \frac{SS_A}{k}$$
$$S_E^2 = \frac{SS_E}{n - k - 1}$$



単回帰について

分散分析をしています。

単回帰分析 分散分析 続き

kS_A^2/σ^2 が自由度 k のカイ二乗分布、 $(n-k-1)S_E^2/\sigma^2$ が自由度 $n-k-1$ のカイ二乗分布に従うことから

$$F = \frac{S_A^2}{S_E^2}$$

は自由度 $k, n-k-1$ の F 分布に従うようになります。検定は片側検定で計算された F 値より自由度 $k, n-k-1$ の F 分布で大きな値をとる確率がこの場合の p 値になることを用いて行います。次の様な分散分析表を書きます。

	自由度	平方和	不偏分散	F 値	p 値
モデル	k	SS_A	S_A^2	$F = S_A^2/S_E^2$	$Prob(X > F)$
誤差	$n-k-1$	SS_E	S_E^2		
総変動	$n-1$	SS_T			



単回帰について

分散分析をしています。

日射量、温度の例で残差平方和などを計算すると次の様になる。

```
# 残差平方和などの計算
> (summary01.SSE <- sum(summary01.resid^2))
[1] 12.06467
> (summary01.SST <- var(summary01.df$temp)*15)
[1] 23.4603
> (summary01.SSA <- summary01.SST-summary01.SSE)
[1] 11.39562
> (summary01.SE2 <- summary01.SSE/(16-1-1))
[1] 0.8617625
> (summary01.SA2 <- summary01.SSA/1)
[1] 11.39562
# F 値
> (summary01.F <- summary01.SA2/ summary01.SE2)
[1] 13.22362
# p 値
> 1-pf(summary01.F,1,14)
[1] 0.002696432
```

この例題の場合は次のようになり、回帰分析することが意味を持つことになる。

	自由度	平方和	不偏分散	F 値	p 値
モデル	1	11.39562	11.39562	13.22362	0.002696432
誤差	14	12.06467	0.8617625		
総変動	15	23.4603			