

IRデータ分析演習

高知大学 大学教育創造センター

高畑 貴志

takashi-takabatake@kochi-u.ac.jp

1

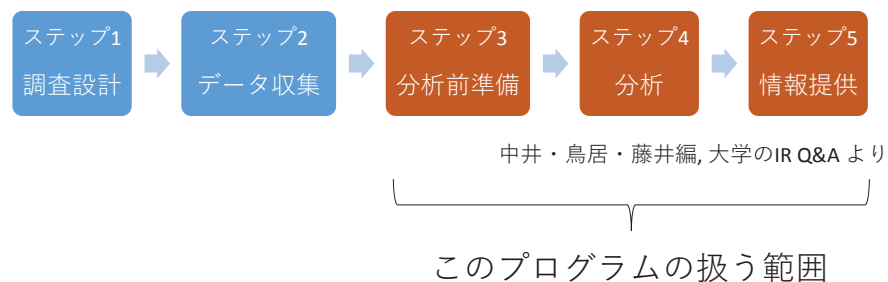
IRとは

IR (Institutional Research) とは、

「高等教育機関の計画、政策形成、意思決定を支援する情報を提供するために機関内で行われる調査研究」(Saupe 1990)

2

IR実践5つのステップ



[分析前準備]: データのチェックと修正、データの結合

[分析]: 基礎的な統計量の算出、

ヒストグラム・度数分布・円グラフの作成

クロス集計、散布図と相関係数による分析

[情報提供]: 効率的な属性（学部・学年等）別のグラフ作成

3

本プログラムの到達目標

1. 初歩的なデータ処理と分析を、資料を参考にして、Excelで行える。
2. キーワードに挙げたExcelの仕組みを、資料を参考にして、IRの実務で活用できる。

(キーワード)

絶対参照／相対参照, COUNTIF関数, VLOOKUP関数, フィルタ機能
ピボットテーブル, グラフテンプレート, リンク貼り付け

4

本プログラムの進め方

導入, 分析, 分析前準備, 情報提供 の4パート構成です。

分析, 分析前準備

- 最初に簡単な説明をします。
- 用意された課題を説明資料を参考に仕上げてください。
- 困ったら、講師や補助の先生に質問してください。
- 余裕があれば、練習問題もやりましょう。
- 早く仕上がったら(1)他の人の補助、(2)説明資料の[発展]を見る、(3)先の課題に取り掛かる、等してください。

情報提供

- 説明が中心となります。

グループ内でのピア・ティーチングをお願いします！

5

準備した模擬データ

「01作業データ」というExcelの「データ」シートの中身

3学部で構成される大学の、ある年度の1～4年生に関する3種類のデータを結合したもの

(実際のデータの傾向に似せて作成した架空のデータ)

- (学籍) 基本情報(学年, 学部, 性別, 入試種別等)
(A～G列)
- (成績) 1年から現在までの平均点、修得単位数等
(H～O列)
- (アンケート) 授業への取り組み、授業外学修時間、身に付いた能力、就職への意識、大学教育の満足度など（設問・選択肢は「設問」シートに記載）
(P～AU列)

6

Part1. 分析

7

分析パートの内容

Excelファイル「01作業データ」を使用

1. 分析前の下準備
2. 基礎的な統計量の算出
3. ヒストグラムの作成
4. アンケートの集計（円グラフ・度数分布）
5. クロス集計による分析

Excelファイル「02散布図用」を使用

6. 散布図と相関係数による分析

8

1. 分析前の下準備

課題

- 「データ」シートの複製を作り「データ.元」という名前にする。以降の操作は「データ」シートで行うこと。
- 学籍番号変換値の左に列を挿入し、1行目はNoという見出しをつけ、2行目以降に、1～1540と連番を振る。
- スクロールしても1行目とA, B列が常に表示されるよう、「ウィンドウ枠の固定」を行う。
- 1行目に「縮小して全体を表示する」の設定を適用する。

説明 (p.2～)

「データ.完成」シートに完成状態を示してある。

9

2. 基本的な統計量の算出

課題

「1.分析前の下準備」の課題を完了した「データ」（もしくは「データ.完成」）シートで、I～P列のデータ(2～1540行目)について、以下の基本的統計量を I1543:P1550の範囲に求める。

最大値、最小値、平均値、分散、標準偏差
中央値、最頻値、データ数

説明 (p.7～)

10

データの尺度

- 数値で表現するデータは、その実質に応じて、次の4つの尺度で分類できます。下の尺度は、上の尺度の持つ性質を兼ね備えます。

•名義尺度

- 1：野菜、2：魚、3：肉等、区別できることに意義がある。
- あ：野菜、い：魚、う：肉、としても意味は変わらない。

•順序尺度

- 1位、2位、3位のような順番の数値。大小は意味を持つが、数字の差は意味を持たない。

•間隔尺度

- 日付、時刻、（絶対温度以外の）温度のように差が意味を持つが、数値の比、n倍といった比べ方に意味がないもの。

•比例尺度（比率尺度とも）

- 金額、距離、時間（時刻ではない）、重量等のように数値の比にも意味があるデータ。最も自然な数値データ。

質的データ

量的データ

尺度によって適用する分析が異なってきます。

11

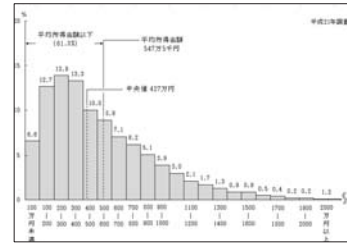
要約統計量とは

- 分布の特徴を表す値（統計量）。記述統計量とも。分布は複数のデータから構成されるが、これらのデータから一定の方式により導き出される「統計量」という単独の値で代表させることで、分布の把握や比較を容易にできる。
- 分布について知りたい側面に応じて、いくつかの統計量が用いられる。
- 平均的な値：平均値(mean), 中央値(median), 最頻値(mode)
- 散らばり具合：標準偏差(standard deviation), 分散(variance), 最大値(max)・最小値(min)

12

平均的な値を表す統計量

- 平均値(mean, average)
 - 全データの和÷データ数
 - 量的データで意味を持つ。



平成21年国民生活基礎調査の概況より

- 中央値(median)
 - データを小さい順に並べたとき真ん中に来る値。データが奇数個であれば、中央の2つの値の平均とする。
 - 平均値に比べて、外れ値の影響を受けにくい。
 - 順序尺度以上で意味を持つ
- 最頻値(mode)
 - 最も沢山現れた値。
 - 名義尺度でも利用可能。量的データにはそぐわない。

散らばり具合を表す統計量

- 分散(variance)
 - 「各データの値－平均値」の2乗の平均値（標本分散）

$$\frac{(\text{値1} - \text{平均})^2 + (\text{値2} - \text{平均})^2 + \dots + (\text{最後の値} - \text{平均})^2}{\text{データ数}}$$
- 推測統計の分野では「データ数」でなく「データ数－1」で割る定義（不偏分散）が使われる。
- 標準偏差(standard deviation)
 - 分散の平方根となる値。
 - データの値を x 倍にすると、分散は x²倍になるが、標準偏差はxになる。
- 範囲(range) = 最大値－最小値
 - 順序尺度以上で利用可能。
 - 外れ値の影響を受けやすい

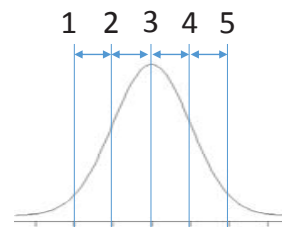
量的データに用いる

[補足] 多段階評価での平均

アンケート調査などで、程度に応じて、

5. そう思う
4. ある程度そう思う
3. どちらともいえない
2. あまりそう思わない
1. そう思わない

等と問う設問への回答(1～5)は、厳密には順序尺度であるが、選択肢間を等間隔であるとみなして、平均などを求めることも多い。

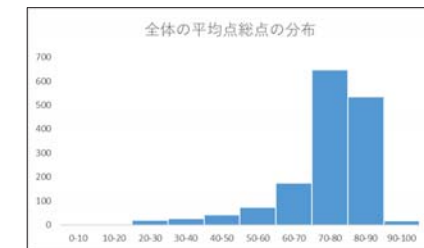


3. ヒストグラムの作成

課題

「データ」（もしくは「データ.完成」）シートの、平均点総計のヒストグラムを、0～100までを10の階級に区切って作成。

完成例



練習問題

修得単位数のヒストグラムを作成。

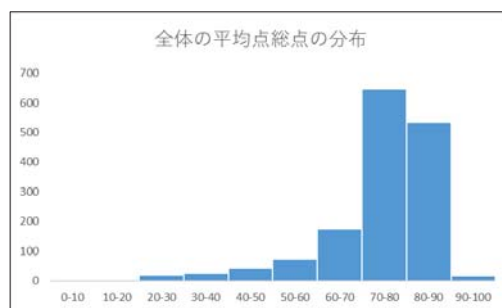
（階級の設定は、各自で適宜に設定してよい）

説明 (p.9～)

ヒストグラム

連続して変化する数値属性について、数値の上下限による「階級」を定め、該当するデータ数（度数）を集計したグラフをヒストグラムと呼びます。

ヒストグラムで分布の形状を確認することは、データ分析の基礎となります。



17

4. アンケートの集計

課題 (次スライドに完成例)

- 択一式の多肢選択問題 **Q-1-1**(真面目に授業に取り組む)を回答した**785**名の内訳を%付の円グラフにまとめる。
- 複数選択可の **Q3** (卒業時まで力を入れたいこと)の回答を、選択肢a~hについて、度数分布グラフにまとめる。

練習問題

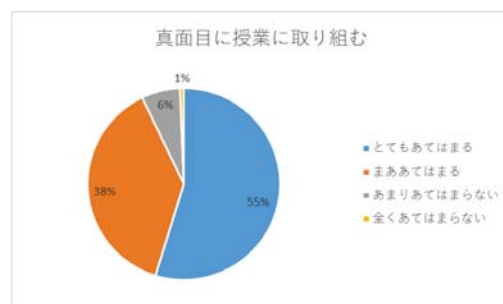
- 択一式の多肢選択問題 **Q-7-1**(大学教育の満足度)を回答した**785**名の内訳を%付の円グラフにまとめる。
- 複数選択可の **Q5** (やる気を向上させたこと)の回答を、選択肢a~cについて、度数分布グラフにまとめる。

説明 (p.24~)

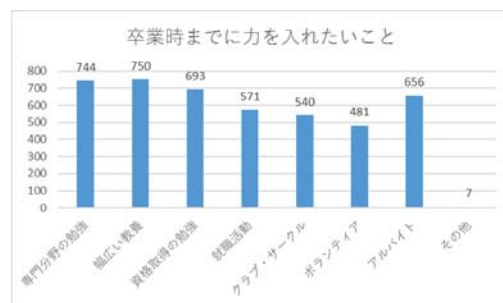
18

課題の完成例

Q-1-1の円グラフ



Q3の度数分布グラフ



19

集計するアンケートの一部

問1 あなたは、大学の授業にどのように取り組んでいますか。

(1), (2)について、最も当てはまる選択肢の一つを選び、選択肢を丸で囲んでください。

(1) 真面目に授業に取り組む

- とてもあてはまる
- まああてはまる
- あまりあてはまらない
- 全くあてはまらない

問3 あなたが卒業までに力を入れたいことを、教えてください。

a~h で該当するものを、記号を丸で囲っていくつでも選択してください。

- a 専門分野の勉強
- b 幅広い教養
- c 資格取得の勉強
- d 就職活動
- e クラブ・サークル
- f ボランティア
- g アルバイト
- h その他 ()

20

アンケートの回答をExcelにまとめた例

通し
番号

択一 (Q1-1,Q1-2,Q2)

複数選択(Q3 a~h)

	A	B	Q	R	S	T	U	V	W	X	Y	Z	AA
1	No	学籍番号変換値	Q-1-1	Q-1-2	Q-2	Q-3-a	Q-3-b	Q-3-c	Q-3-d	Q-3-e	Q-3-f	Q-3-g	Q-3-h
2	2	kb86c15dd	3	3	2		1	1	1				
3	3	k2c0c932d	4	4	3	1	1	1					
4	5	k189dbb8d	4	4	9	1	1		1				語学
5	9	k0408b34a	3	3	4	1	1	1	1	1	1	1	卒研
6	11	k93ebd1b8	4	3	3	1	1	1	1		1	1	ゼミ、公
7	12	kd5e68e60	3	3	3	1	1	1		1	1	1	教員採

手際よく集計するためのポイント **Point!**

- 通し番号をふる。
- 1人1行に収める。
- 択一の回答は1問1列
- 複数選択可能な選択問題は、選択肢ごとに列を用意し、選択されたものにのみ"1"を記入
- その他 () には1列用意
- 自由記述は、複数行になっていても1セルにまとめる。

21

5. クロス集計による分析

課題 (次スライドに完成例)

- 男女別にQ-1-1(真面目に授業に取り組む)の回答数を集計
- 学年別・学科別に修得単位数の平均を集計

練習問題

- Q-1-1(真面目に授業に取り組む)の回答とQ-7-1(大学教育の満足度)をクロス集計
- 学年別・学科別に平均点総計の平均を集計

説明(p.36~)

22

課題の完成例

必要最小限の状態です。
適宜、見栄えの修正を!

男女別のQ-1-1の回答

データの個数 / 学籍番号変換値	列ラベル	1	2	3	4	総計
行ラベル						
女		2	19	131	206	358
男		3	32	168	224	427
総計		5	51	299	430	785

学部・学年別の修得単位数の平均

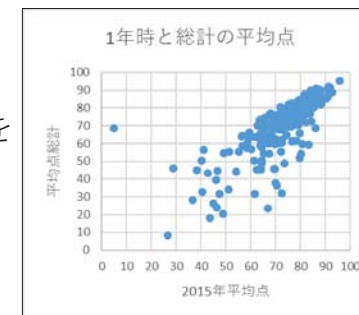
平均 / 修得	列ラベル	A	B	C	総計
行ラベル					
1		20.2	20.4	21.0	20.5
2		62.0	64.2	69.8	65.1
3		105.3	104.5	121.4	110.1
4		119.9	117.0	151.0	128.1
総計		79.3	78.7	91.3	82.8

23

6. 散布図と相関係数による分析

課題

- Excelファイル「02散布図用」の「4年」シートの平均点2015と平均点総計の関係を表す散布図を作成し、その相関係数を求める。



練習問題

- 「4年A学部」シートの平均点2015と平均点総計の散布図と相関係数を求める。

説明 (p.43~)

24

散布図

- 数値データとなる二種類の属性の関係を、一方の値を横軸に、もう一方の値を縦軸に取ってプロットしたグラフを散布図という。



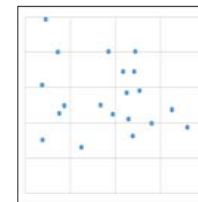
- 散布図により2つの量の間の関係を見ることができる。

25

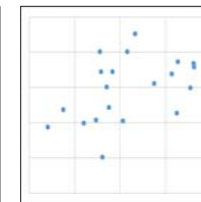
相関係数

- 散布図により示される二つの属性の関係が、直線関係にどれだけ近いかを示す指標に「相関係数」が用いられる。相関係数は r で表記する。
- 相関係数の絶対値が1に近いほど直線に近い関係
- 相関係数がプラスなら右上がり、マイナスなら右下がりの関係

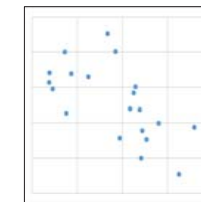
$r=-0.20$



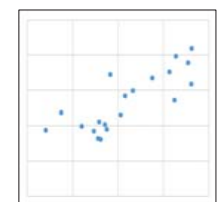
$r=0.45$



$r=-0.60$



$r=0.80$



26

相関係数による相関の評価

- 相関係数の大きさに2項目間の相関の強さを評価できる。
- 評価の基準の例を以下に示す。

$0 \leq r < 0.2$	ほとんど相関なし
$0.2 \leq r < 0.4$	弱い相関あり
$0.4 \leq r < 0.7$	中程度の相関あり
$0.7 < r \leq 1$	強い相関あり

評価の基準は研究分野等により、多様なものが使われているので、注意すること。

27

相関係数に関する注意点

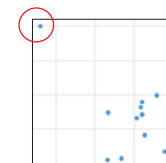
- 相関係数は、「直線の関係」への当てはまりを見るもの。直線でない関係は扱えない。(円状の右図の場合、相関係数=0)



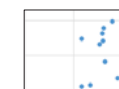
- 相関があっても、因果関係があるとは限らない。
- 第3の要因のため、直接関係のない2つの量に相関がみられ場合がある。(例:小学生の体重と学力)

- 相関係数は、外れ値の影響を強く受けることがある。

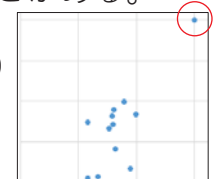
$r=-0.51$



$r=0.37$



$r=0.79$



28

Part2. 分析前準備

29

分析前準備パートの内容

1. データのチェック
Excelファイル：03データチェック
2. データの結合
Excelファイル
04a学籍, 04b成績, 04cアンケート
3. データの抽出
Excelファイル：05データ抽出用

30

1. データのチェック

課題

「03データのチェック」には、4段階の多肢選択問題Q-1～Q-5への50人分の回答がまとめられている。

適切な選択肢は1～4で、無回答は空欄にすることになっているが、以下のように適切でない箇所が含まれる。

- "5" がある。
- 未回答の空白の代わりに0や-と記入してある
- 2.3 や 1,3 などのセルがある
- 空白ではなく、全角や半角のスペースが含まれる

[タスク] 不適切な箇所を見つけ、空白に置き換える。

説明 (p.47～)

31

その他のチェック

- 数値データでは、最大値、最小値、ヒストグラムによる分布の状況などを見ることで、異常な値をチェックすることができます。
- 多肢選択問題で全て同じ選択肢を選択している場合等は、設問数が多ければ分析から除外するのが適切です。

今回は不適切な値は「無回答」に置き換えましたが、別の方針で処理する場合があります。定められた方針（なければ定めて）に従うようにしてください。

32

2. データの結合

課題

- 「04a学籍」の表の右に、「04b成績」の表から「学籍番号変換値」が同じ行のデータをつなげる。
(表の結合)

練習問題

- さらに「04cアンケート」の「回答」シートのデータを、つなげる。

説明

PowerQueryを用いた場合 (Excel2016以降) p.60～

PowerQueryを用いない場合 p.70～

33

3. データの抽出

課題

- 「05データ抽出用」のデータを用い、以下の操作を行う。
 - B学部のデータを抽出
 - Q-1で3または4と回答した学生を抽出
 - Q-1で3または4と回答した女子学生を抽出
 - 平均点総計が60点未満の学生を抽出
 - 2014年度入学生を削除したデータセットを作成

フィルター機能の使用を想定しています。

a～d は、該当するデータのみが表示された状態にすればOKです。

説明 (p.82～)

34

Part3. 情報提供

35

情報提供パートの内容

Excelファイル「06a学部別レポート」

Wordファイル「06b学部別レポート」
を使用

- COUNTIFによる択一式多肢選択問題の一括集計 (p.91～)
- グラフテンプレートの利用 (p.94～)
- 学部別の自動集計 (p.97～)
- リンク貼り付けの活用 (p.100～)

詳細は説明 (p.80～) をご覧ください。

36