(4)「情報通信ネットワークとデータの活用」で学ぶこと

- 要するに
  - インターネットの仕組み
  - 暗号
  - オープンデータ
  - データベース(Access)
  - 量的データの分析(Excel)
  - 質的データの分析
  - データの可視化

学習22, 23, 24

学習22 量的データの分析

(1)量的データと質的データ

(2)量的データ間の関係

(3)単回帰分析を用いた値の推測

(4)量的データの統計的仮説検定

(1) 量的データと質的データ

- 構造化データ
  - 行が観測対象、列が観測項目に対応した表形式のデータ
  - 非構造化データの例：文書・音・画像など
- データの種類
  - 時系列データ：時点が対象となり、その系列に沿って得られるデータ
  - クロスセクションデータ：時点が固定され、同質の対象の複数観測結果からなるデータ

(1) 量的データと質的データ

- データ項目の種類と尺度
  - 量的データ：数直線上の値として得られるデータ
    - 間隔尺度：気温など、絶対的な原点がなくデータの差は意味を持ち、データの比は意味を持たないもの
    - 比例尺度：身長など、絶対的な0を持ち、データの差も比も意味を持つもの
  - 質的データ：分類項目の対象を表すデータ
    - 名義尺度：男、女など、対象についての分類
    - 順序尺度：優良可など、順序構造を持つ分類

(1) 量的データと質的データ

- データ項目の種類と尺度

種類	尺度	例	大小比較	差	比
質的	名義尺度	学生番号	×	×	×
質的	順序尺度	成績の順位	○	×	×
量的	間隔尺度	気温	○	○	×
量的	比例尺度	重さ	○	○	○

(1) 量的データと質的データ

- データの変容
  - 欠測値の処理
    - 欠測コードを決め、通常のデータと区別できるようにする
    - 欠測値を含む行を削除するか、欠測値を近いと思われる値で補間する
      - ✓ 補間方法の例：平均値、中央値

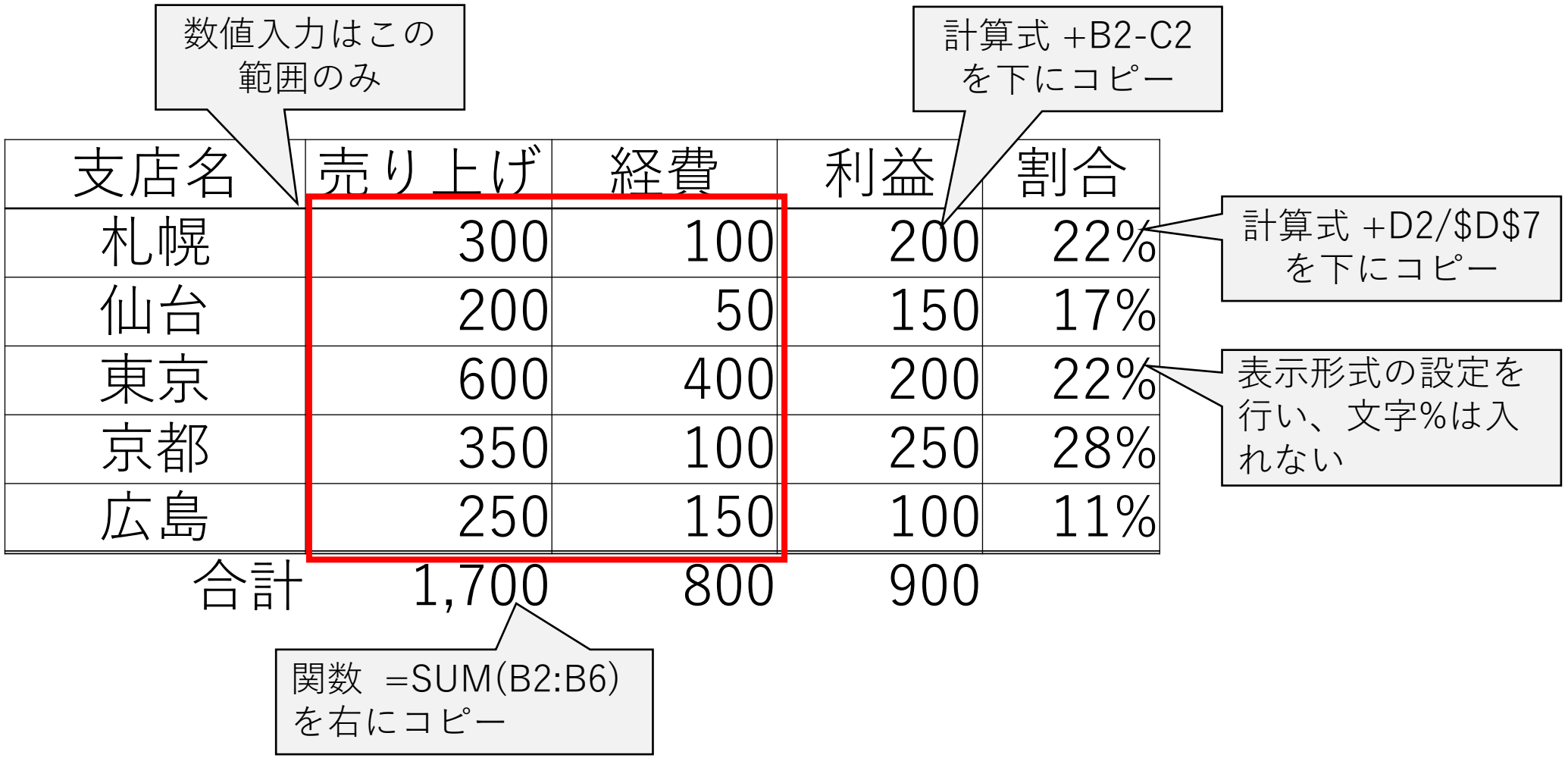
(1) 量的データと質的データ

- データの分析
  - 層別
    - 質的データの値を基にデータをグループ分けして比較分析すること
      - ✓ 例:地域別、男女別
  - 変数変換
    - 規模等の影響を取り除くために、適当な単位あたりの数に変換する
      - ✓ 例:人口10万人当たり
  - 水準化
    - 量的データをいくつかに区切って順序尺度に変換する
      - ✓ 例:年代別
  - 統計的分析
    - 平均値・中央値・最頻値や、四分位数・最小値・最大値を箱ひげ図で表す

補足：表計算ソフトExcel

- 表作成
  - 極力間違いが入らない、検証可能な表を作る
  - ソートやフィルタの機能も有効に活用する
- グラフ作成
  - 対象とするデータの特性に合わせた適切なグラフを作る
- データベース
  - 簡易なデータベース機能

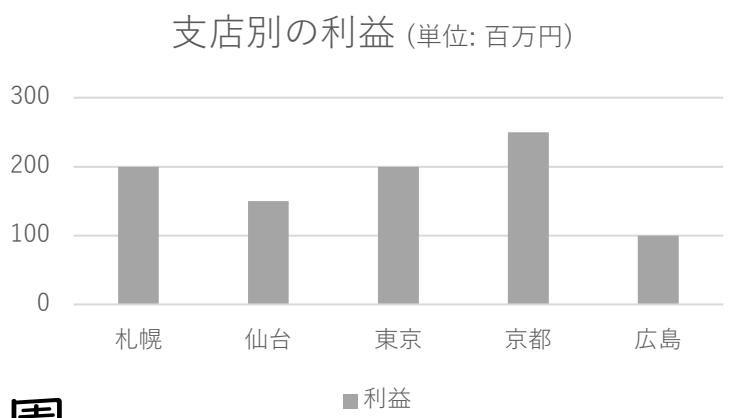
補足：表計算ソフトExcel 表作成

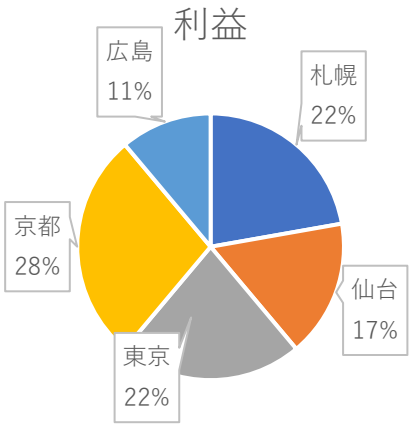


支店名	売り上げ	経費	利益	割合
札幌	300	100	200	22%
仙台	200	50	150	17%
東京	600	400	200	22%
京都	350	100	250	28%
広島	250	150	100	11%
合計	1,700	800	900	

補足：表計算ソフトExcel グラフ

- グラフにおいて明記するもの
  - グラフ名、縦軸・横軸があらわすもの、凡例、単位
- グラフの種類
  - 棒グラフ、円グラフなど
  - 使い分けの参考資料
    - 総務省統計局 なるほど統計学園
      - ✓ 4. グラフの作り方（初級編）http://www.stat.go.jp/naruhodo/4_graph/index.html
    - 代表的な4つのグラフの使い分けのポイント
      - ✓ https://webtan.impress.co.jp/e/2008/10/29/4281





補足：表計算ソフトExcel データベース

- データベース
  - 小さな表をデータベースとして扱う

=XLOOKUP(検索値、検索範囲、戻り範囲)

検索値

書籍ID	書籍名	作者ID	作者名
1	羅生門	1001	芥川龍之介
2	こころ	1002	夏目漱石
3	坊ちゃん	1002	夏目漱石
4	鼻	1001	芥川龍之介
5	舞姫	1003	森鴎外
6	吾輩は猫である	1002	夏目漱石

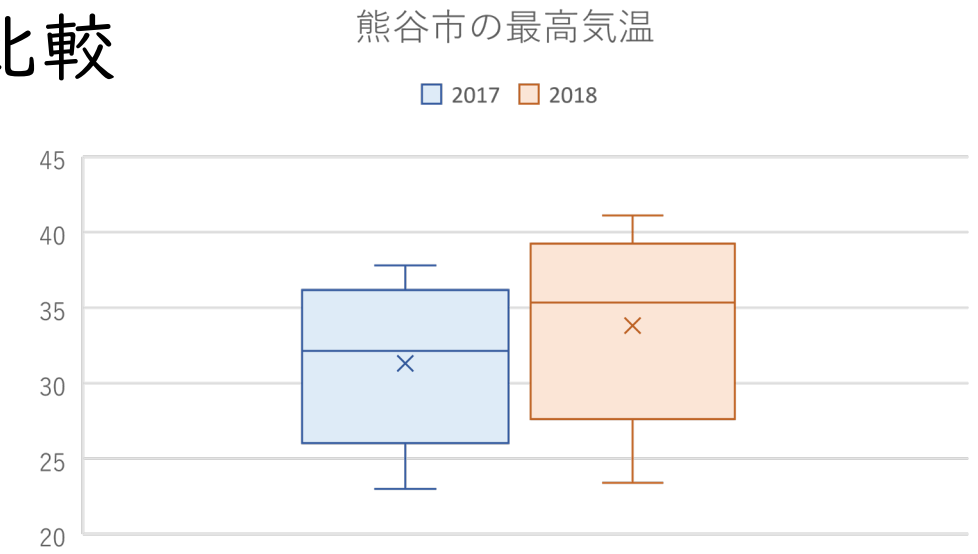
作者ID	作者名
1001	芥川龍之介
1002	夏目漱石
1003	森鴎外

検索範囲　　戻り範囲

(1) 量的データと質的データ

- 演習1
  - 2017年と2018年の埼玉県熊谷市の7,8月の最高気温のデータを比較して、2018年が本当に暑かったのかを検証

1. 気象庁のサイトから当該データをダウンロード
   - http://www.data.jma.go.jp/gmd/risk/obsdl/index.php
2. Excelでグラフを作成して比較
   - 箱ひげ図
     - ✓ 上の線から順に
       最大値、第3四分位、中央値、
       第1四分位、最小値
     - ✓ ×は平均値

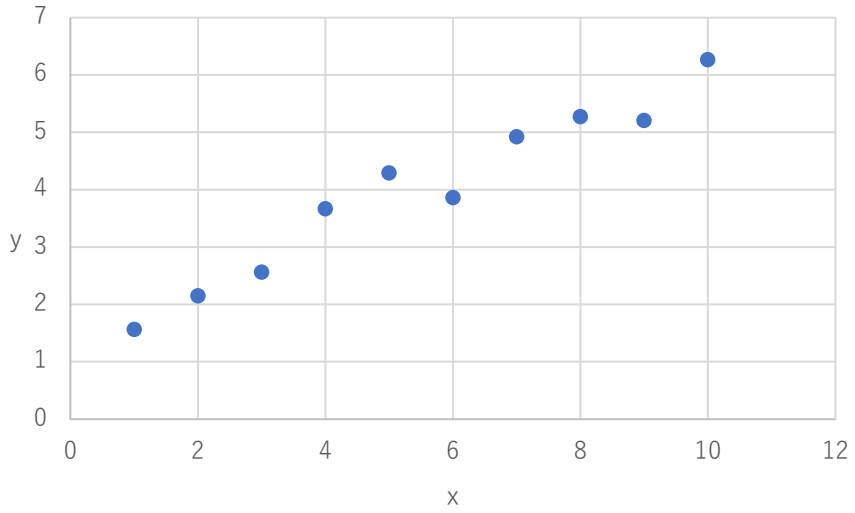


(2) 量的データ間の関係

- 量的データ間の関係
  - 2つの量的データx, yの関係は、散布図によって傾向を捉え、相関係数によって定量的に相関関係を捉えることができる

相関係数　$r = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$

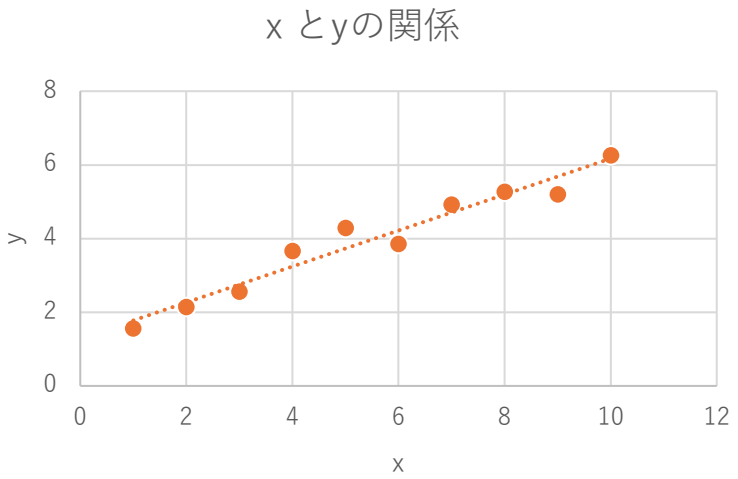
$\bar{x}$　平均値　　σ_x　標準偏差



相関係数
- 0<r≦1　正の相関
- 0　無相関
- -1≦r<0　負の相関

(3) 単回帰分析を用いた値の推測

- 単回帰分析とは
  - x, yの関係を $y=\alpha x+\beta$ で表し、傾きや切片を求めること
- 単回帰分析の方法
  - 直線上の値と実際のyの値との差を残差とよび、残差を2乗した値の総和が最小になるような回帰直線の決定法を最小二乗法とよぶ
- Excelでの分析
  - 回帰直線を引くだけ
    - 「ホーム」→「データ分析」→グラフを選択
  - 回帰直線を求める
    - 傾き：SLOPE関数、切片：INTERCEPT関数



(4) 量的データの統計的仮説検定

- 統計的推測とは
  - 未知の母数（母集団の特性値）を標本データから推測する方法
  - 点推定と区間推定がある
- 仮説検定とは
  - 母数に関する仮説の真偽を検証すること
  - 検証したい仮説を対立仮説、その反対の仮説を帰無仮説とよぶ
  - 標本データが帰無仮説上でどのような分布の下で起こっているかを評価し、その確率（p値）が小さければ帰無仮説を棄却し、対立仮説の成立を主張する

(4) 量的データの統計的仮説検定

- Excelで仮説検定
  - アドイン「分析ツール」をアクティブに設定
  - 「データ」→「データ分析」→「t検定：等分散を仮定した2標本による検定」
    - 範囲はラベル含んで指定し、「ラベル」にチェックを入れる

学習23 質的データの分析

(1)質的データの種類とその扱い

(2)テキストデータの扱いについて

(3)テキストデータの可視化

(4)テキストの分析とその可能性

(1) 質的データの種類とその扱い

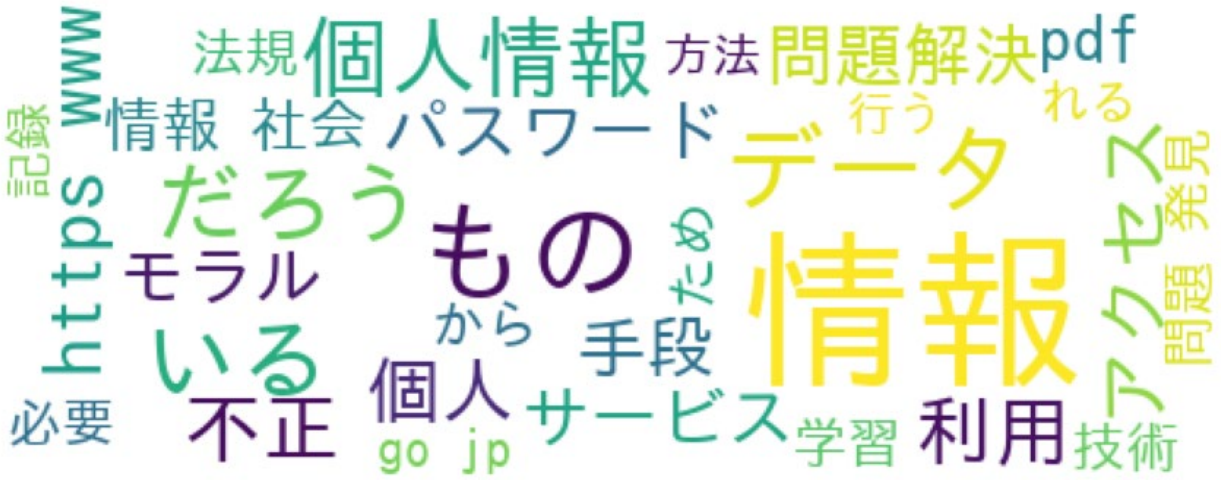
- 順序尺度を扱うときの注意
  - 数値で表されているときでも、平均などの集計が意味を持たない場合もある
    - 順序にしか意味がない場合：学年など
    - 順序にすら意味がない場合：市町村コードなど
  - 主観評価の場合
    - 回答形式によって回答者の意識が変わることがあることに注意
      - ✓ 例：リッカート尺度（文を提示してそれに対して回答する）と数値評価の場合

(2) テキストデータの扱いについて

- テキストデータの扱い
  - 従来は、手作業による分類しかできなかった
  - 近年では、自然言語処理技術によってアンケート等の自由記述を自動的に分析できるようになった
    - 日本語の書き言葉では単語の区切りが明示されないので、単語分割の処理が必要になる

(3) テキストデータの可視化

- テキストマイニング
  - テキストマイニングツールは、日本語の文章を入力すると、単語の出現頻度をフォントの大きさで示すワードクラウドや、出現頻度統計などを計算して表示することができる



ワードクラウドの例

(4) テキストデータ分析の応用

- 応用分野
  - 言語学的分析
  - SNSの発言分析
  - スマートスピーカーによる発話分類

  など

学習24 データの形式と可視化

(1)質的データとその種類

(2)データの分析と可視化

(3)データの可視化と問題発見

(1) 質的データとその種類

- データの種類の組み合わせによる可視化法の違い
  - 例) 男女別の身長データ
    - x: 名義尺度、y: 比例尺度
    - 箱ひげ図や分布の様子を表すヴァイオリンプロットで表現
  - 例) 5段階評定により英語と数学の成績の比較
    - x: 順序尺度、y: 順序尺度
    - 分割表 (クロス集計表) やマリメッコチャートで表現

(2) データの分析と可視化

- 表形式ではないデータの可視化
  - 多対多の関係：有向グラフ
- 表形式のデータ
  - BIツールの活用
    - 質的データと量的データを自動判別して集計、可視化が可能

(3) データの可視化と問題発見

- データの可視化による問題発見
  - 可視化結果が複雑な場合に分析が必要
  - 対象としている2変数と相互に相関の高い隠れた変数である交絡因子を考慮する必要もある

補足：テキストマイニング

- テキストマイニングの手順（ワードクラウドを例に）
  - 文章の単語分割
  - ストップワードのリストアップ
  - ワードクラウドの表示
    - フォントの大きさは単語の出現回数に依存
    - 似ている出現回数の単語を同じ色に。ただし、見やすくするために色の系統を制限しているので、頻度が異なっても同じような色に見えることがある

参考文献・資料

- データ可視化など
  - 北川他：教養としてのデータサイエンス、講談社、2021.

大学入試センター サンプル問題『情報』第3問 問1 (p.14)

- 解答
  - a
    - ア 0 得点
    - イ 3 反則回数
    - ウ 3 D
    - エ 2 全参加チームについて正の相関がある項目の組合せの中には，決勝進出チーム，予選敗退チームのいずれも負の相関となっているものがある。
      （シンプソンのパラドックス）