R で 6 週間学ぶ新しい統計基礎

最終更新 2021 年 4 月 16 日版

2021 年 4 月 5 日作成

今回のワークショップの動機

この 6 週間の目標とデザイン

言語学のデータと統計

応答変数の分布：確率変数・確率分布の導入として

今回のワークショップの動機

今回のワークショップの動機

統計は，言語学者の議論の対象外なのか？

- 言語学者が言語現象を数値で表現する際の，その数値表現の捉え方は，言語学者が議論する対象ではないのだろうか
  - 「数の話はどうでもいい，言語学の話をしろ」

統計の威力とその限界はどこにある？

- 統計は誰がやっても同じ結果を導くのだろうか？
  - 「統計を掛けるマニュアル・プログラミング方法」に従っておけば，どんなデータも，その方法ひとつで調べられる？

今回のワークショップで強調したいこと

統計を使う限り，統計も，言語学者の議論の対象である

- 言語学者が言語現象を数値で表現する時には，その数値表現にも，言語学者が手厚く議論すべきでは？
  - 「数を証拠にする以上，その数の扱いも問われるのでは？」

今回のワークショップで強調したいこと

あるデータに対する統計は，複数考えられる

- ここでの統計：**推測統計**（単なるデータの要約ではない）
- 統計の掛け方で，あるデータに対するその人のデータの見方が分かる場合がある
- 言語学の議論と同様に，統計の掛け方にも，言語学者の独創性が現れる
- 手許のデータに対する統計は，あらゆる多様な分析の中の 1 つに過ぎない
  - だからこそ，なぜその統計の掛け方をしたのか説明する必要がある

記述統計ではなく，推測統計がしたい

- 手許のデータ（標本のひとつ）から，手許のデータを作る根本の原理（母集団）を知りたい
- 個々の標本も大切で重要だが，その背景にどのような原理があるのか知りたかったのでは？

記述統計と推測統計

記述統計

- **得られたデータ（観測値）そのもの**について，「データの表す集団の性質を記述し要約する」(日本統計学会, 2017, p.55)
  - **標本集団**

推測統計

- 「そのデータの基になっている集団について推測する」(日本統計学会, 2017, p.55)
  - **母集団**
- 「データを取る段階から，得られたデータを分析し推理・推論するまでの方法論全体」
  1. 母集団からの無作為抽出
     - **1.1** 標本調査法
     - **1.2** 実験計画法
  2. 標本から平均や分散を計算し，母集団について推論する
     - **2.1** **統計的推測**

この，推測統計の統計的推測がこのワークショップのテーマ

標本の記述と母集団の推測の関係

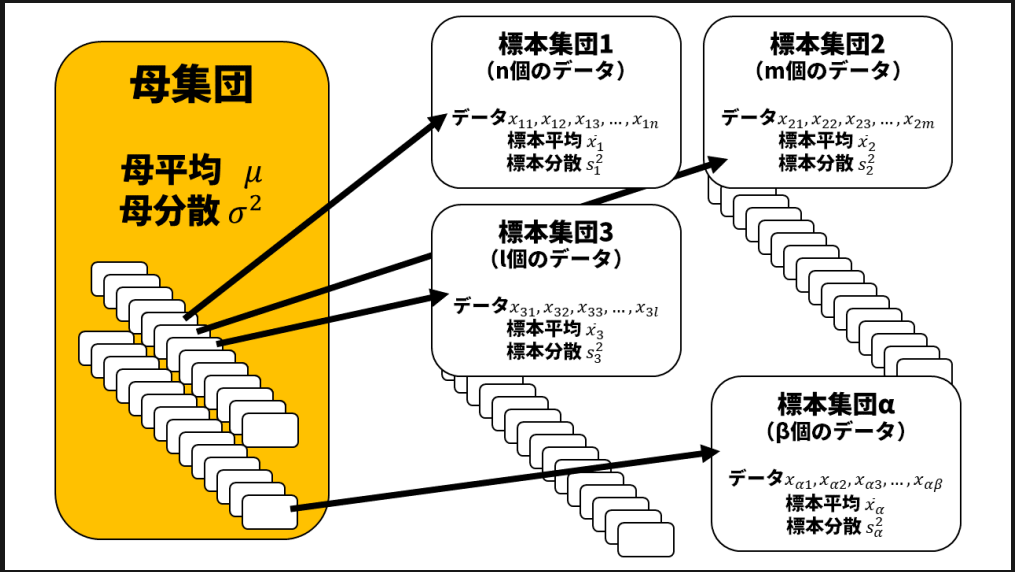


Figure 1: 母集団と標本集団の関係

多様な統計の世界：統計の理論とその応用

統計の理論

- 1970 年代には，既に**様々な性質を持つデータに対応しうる統計理論のひとつ**が出来ていた
  - 一般化線形モデル（1972）

統計の運用

- 2000 年代以前までは，コンピュータの制約上，データは多様なのに，決まった統計の方法しか出来なかった
  - 「誰がやっても同じ」= それしか方法がない
  - t 検定（理論は 1870 年代より）
  - 分散分析（理論は 1920 年頃より）
- 2000 年代以降，一般のコンピュータで，その人のデータの実態にあった統計が掛けられるようになった
- にもかかわらず，言語学では（言語学だけではないけれど），いまだに，データの実態から乖離した「誰がやっても同じ」統計をかけ続けている例がある
- **データの実態にあった統計が掛けられる手段はあるのだから，それを自ら選んで使ってみないか**

データの実態にあった統計が掛けられるようになったからこそ

その独創的な統計を「誰がやっても同じ結果になる」ようにする（**再現可能性**を持たせる）には，相当な前提知識の開示が必要になる

- 「誰がやっても同じ結果になる」= あらゆる手法が考えられる中で，その方法を 1 つだけ選んだ理由を明示し，その方法の手続きを開示する
- **扱う言語現象は何か**
  - どのような数理的振る舞いをすると仮定したか
  - どのように数値で表現したか
- **扱う言語現象に対する要因は何か**
  - どのような数理的振る舞いをすると仮定したか
  - どのように数値で表現したか
- 扱う言語現象と扱う言語現象に対する要因が，どのように数理的に結びついていると仮定したか

この 6 週間の目標とデザイン

今回扱う統計は，「仮説の推測的検証方法」

- **要因に既に心当たりがある場合の統計**
  - **統計的仮説検定**
    - いわゆる「有意差があるか」を調べる統計
  - **統計的な予測**
    - ある要因によって，ある言語現象がどれだけ変動するのか

一般化線形モデルによる仮説検定と予測

以下の式で，統計を掛けられるようになる

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

ちょっと端折って

$$y \sim \beta_0 + \beta_i x_i$$

一般化線形モデルによる仮説検定と予測

モデル式

検証する事象 $\widetilde{y}$ 「近似する」 $\underset{\sim}{\sim}$ $\underbrace{\beta_0}_{\text{事象の基礎}} + \underbrace{\beta_i}_{\text{要因の効果の大きさ}}$ 事象を変化させる要因 $\widetilde{x_i}$

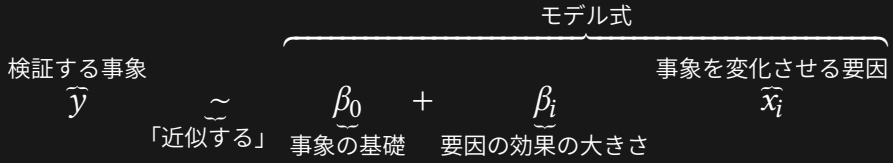
統計的仮説検定

- 要因の効果の大きさが 0 ではないことを証明する
- 帰無仮説：「要因の効果の大きさが 0 である」という主張
- 対立仮説：「要因の効果の大きさが 0 ではない」という主張

予測

- モデル式の中の「事象を変化させる要因」x_i の値が変わると，「検証する事象」y がどのような値を取るのか
- 回帰 (日本統計学会, 2017, pp.32–33)
  - 一方の変数 x から，他方の変数 y を説明する（$x \to y$）
  - **相関**とは異なる
    - 相関：x と y といった変数同士の，相互関連性の強度
    - 相関では，x と y といった変数同士が対等

一般化線形モデルによる仮説検定と予測



今回扱う統計は何ではないか

- 「要因をこれから探る」統計ではない
  - **探索的方法**
    - 主成分分析
    - 対応分析

R でプログラミング

「2000 年代以降，一般のコンピュータで，その人のデータの実態にあった統計が掛けられるようになった」

- 商用ソフトウェアだけでない，無料のソフトウェアの登場
  - R もそのひとつ
    - 個人での開発も盛ん
    - 積極的で公開での開発 ← ユーザからの意見や質問で盛り上がる
    - 案内役の第 1 プログラミング言語

タイムスケジュール案 I

4 月 9 日（金）

- R・RStudio のインストール
- 言語学のデータは数値で表せる？
  - 統計における数値データとは
  - 数値データをどう捉え扱うか
  - 尺度という概念［コーディング］

4 月 16 日（金）

- 仮説検定
  - 「等分散正規分布に従う複数群」の平均値の差を比べる t 検定を例に
  - 23 日と内容を入れ替えるかもしれません

4 月 23 日（金）

- 確率分布
  - 正規分布・二項分布・ポアソン分布を図示しながら考える
  - 16 日と内容を入れ替えるかもしれません

タイムスケジュール案 II

4 月 30 日（金）

- 回帰分析
- モデリング
  - ロジスティック回帰
  - ポアソン回帰
- [尺度とコーディング再び]

5 月 7 日（金）

- 一般化線形混合モデル
  - 個人差をどう扱うか？

5 月 14 日（金）

- まとめと実践
  - 言語学会大会予稿の統計を検討する
  - 事実と乖離する統計になっていないかの検討
  - どのような統計を掛けるのが適切かの検討

各回の進め方 I

- 案内役の説明
  - 案内役は，個別の事例に対するアドバイスには慣れているが，全体像を一人で全部お話しするのは初めて
  - 一方向的な講義になるよりは，一緒に考える時間を持ちたい
  - 数理的な専門家に教わりたい方には，ちょっと向かないかも
- R でシミュレーションデータを自作しながら，統計を掛けてみる
  - シミュレーションすることで，自分がこれから得ようとする本物のデータの性質も理解できるようになる
  - 統計とプログラミングを理解するには，徹底的な産出が必要
    - 語学よりも産出が決定的かもしれない
- ブレークアウトセッションがある
  - 折角ひとつの場を共有するので，コードや統計理論に関して一緒に考えたい
  - 独習が好きな方には，ちょっと向かないかも...

各回の進め方 II

1. 気楽に・気長に聞いてください
   - お茶・お菓子・ご飯のご用意を
   - 統計やプログラミングは，一発で出来るようにならない場合があります
     - 今回のワークショップが入口になって，2 回目・3 回目の勉強につながれば…
     - 今回のワークショップで，1 回目の勉強で分からなかったことがちょっとでも分かるようになれば…
   - エラーが出るかもしれませんが，気長に…
     - 例：
       初回は，ウイルス対策ソフトが，R や RStudio（に関連したソフトウェア）をマルウェアと誤認するかもしれません

各回の進め方 III

2. 説明の途中でも，割って入って質問してください
   - そのための案内役です〜
   - 資料は，後から一人でじっくり読めます
   - 誰かに質問するなら，今がチャンス !!
   - 質問の内容によっては，回答が前後するかもしれませんが，とにかくご質問ください
   - 状況に応じ，作業中の画面を共有していただくようお願いするかもしれません
     - 必要に応じ，「デスクトップアイコンを非表示にする」等を行ってください

各回の進め方 IV

3. ワークショップなので，インタラクティブにできたらと思います！
   - 時間の許す限り，成果物を見せてもらったり，テンプレートをアレンジしてもらうかも
   - こちらから指名する可能性もあります
4. Zoom のお名前を次のように設定してください
   - インタラクティブに参加したい方：
     「（パソコンの OS・Mac/Win）_（氏名）」
     - 例：Win_ 小川雅貴
   - とにかく話を聞くのに集中したい方：
     「聞 _（パソコンの OS・Mac/Win）_（氏名）」
     - 例：聞 _Win_ 小川雅貴

言語学のデータと統計

言語学にはどんなデータがある？

- 2 分以内に最低 5 つ，自分でとにかく考える
  - 出来れば「応答変数（結果）」←「説明変数（要因）」のセットで考えてほしい
  - 手許の端末にタイプしてメモしておく
- 4 人 1 班のブレークアウトセッションで，自分の思いついたデータを共有
  - jamboard が 1 つあり，その中に班ごとのページがある
  - 各班のページに，**黄色の付箋**で記入
  - **黄色の付箋**の内容の内，他の班で挙がってなさそうなデータを 3 つ選び，そのデータを**黄緑色の付箋**に記入
  - jamboard は共用なので，他の班を偵察しないこと！

言語学のデータで数値化できそうなものとできなさそうなものは？

- jamboard のページ（全班共通）に，先ほどの**黄緑色の付箋**をコピーして貼っていく
  - 数値化できる・できないと思った理由は？（**水色の付箋**で）
  - 数値化できなさそうなデータは，無理やりにでも数値化できる？
  - jamboard のコラムに，数値化できそうなデータ・できなさそうなデータの例が埋まらない場合，先ほどの**黄色の付箋**（個々人が思いついた例）もコピーして貼ったり，新規に考えて**黄色の付箋**で貼ったりする

数値化できそうな言語データ

数値化できそうな言語データ

言語学のデータをより細かく分類する

- jamboard のページ（全班共通）に，先ほどの**黄緑色の付箋**（各班で取りまとめた例 3 つ分）をコピーして貼っていく
  - 大分類・小分類には，それぞれどのような特徴がある？
    - 分類の理由・根拠は？この分類より細かい分類もある？（**水色の付箋**で記入）
  - 各分類の欄が埋まらない場合，先ほどの**黄色の付箋**（個々人が思いついた例）もコピーして貼る（新規作成も**黄色の付箋**で）

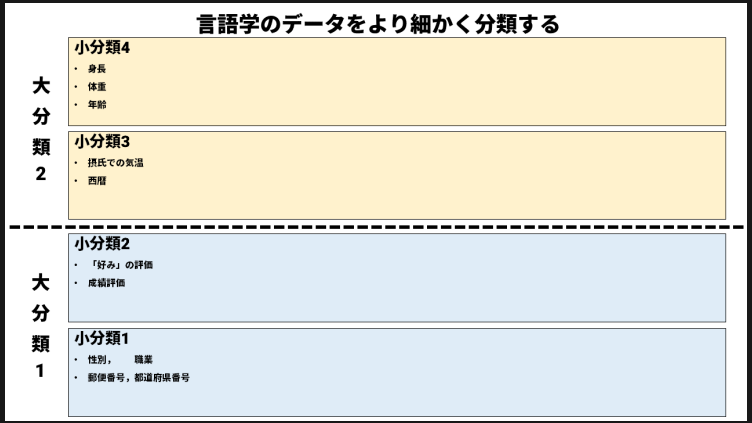


Figure 9: 言語学のデータをより細かく分類する

名義尺度・順序尺度・間隔尺度・比例尺度

大分類	尺度	尺度の意味	要約統計量（「指標」）	例
量的変数（数値）	比例尺度	（一致・不一致） 値の大小関係 差分 比率	度数 最頻値 中央値 平均 標準偏差	身長 体重 年齢 読み時間 基本周波数 F0 単語の産出数
	間隔尺度	（一致・不一致） 値の大小関係 差分	度数 最頻値 中央値 平均 標準偏差	摂氏気温 西暦
質的変数（カテゴリ）	順序尺度	一致・不一致 値の大小関係	度数 最頻値 中央値	「好み」の評価 成績評価 容認度・文法性（3 段階以上）
	名義尺度	一致・不一致	度数 最頻値	性別 職業 郵便番号・都道府県番号

量的変数（間隔・比例尺度）の別分類：離散変数と連続変数

離散変数

- とびとびの値
- **カウントデータ**

連続変数

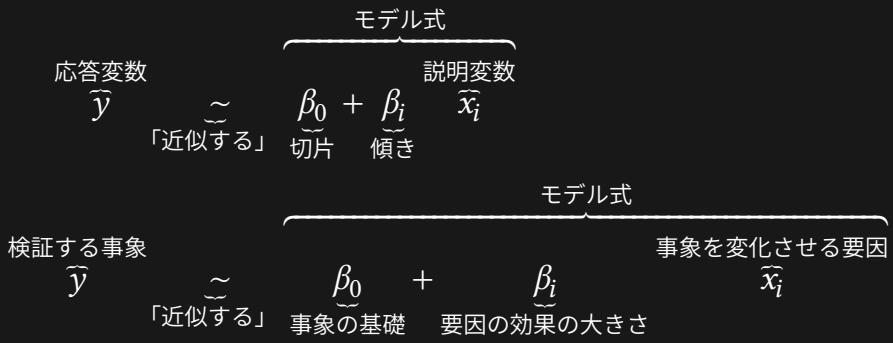
- 「軽量する装置の精度には依存するが，本来は連続量である値」(日本統計学会, 2017, p.6)

名義尺度・順序尺度・間隔尺度・比例尺度／離散・連続変数

大分類	尺度	尺度の意味	要約統計量（「指標」）	例	連続／離散
量的変数（数値）	比例尺度	（一致・不一致） 値の大小関係 差分 比率	度数 最頻値 中央値 平均 標準偏差	身長 体重 年齢 読み時間 基本周波数 F0 単語の産出数	主に連続 主に連続 連続・離散 主に連続 主に連続 離散
	間隔尺度	（一致・不一致） 値の大小関係 差分	度数 最頻値 中央値 平均 標準偏差	摂氏気温 西暦	主に連続 離散
質的変数（カテゴリ）	順序尺度	一致・不一致 値の大小関係	度数 最頻値 中央値	「好み」の評価 成績評価 容認度・文法性（3 段階以上）	（離散的） （離散的） （離散的）
	名義尺度	一致・不一致	度数 最頻値	性別 職業 郵便番号・都道府県番号	（議論に依る） （離散的） （離散的）

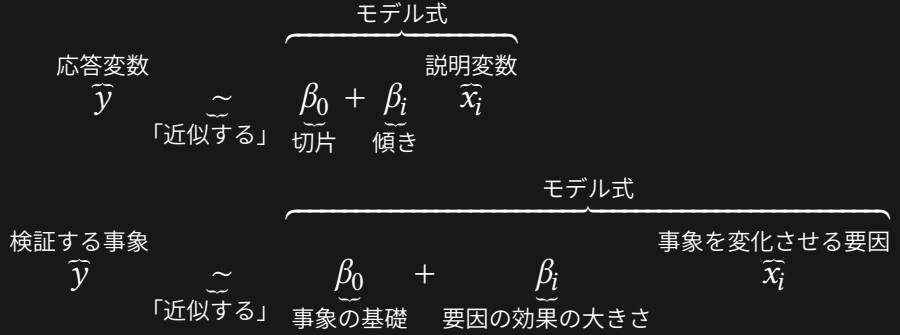
ここまでで押さえてほしい・考えてほしいこと

1. データには多様な尺度があること
2. **4 つの尺度と離散・連続の区別に応じた統計方法**があること
   - **これら多様な尺度に，同一の統計方法は適用できない**
   - **応答変数における 4 つの尺度と離散・連続の区別**に応じた分析法
   - **説明変数における 4 つの尺度と離散・連続の区別**に応じた分析法



ここからの説明の流れ

1. **応答変数における 4 つの尺度と離散・連続の区別**に応じた分析法
   - 応答変数の方が，研究によって取り得る観測値が多様
   - 「観測値の背景 = 母集団」の分布を知る上で，応答変数から先に説明
2. **説明変数における 4 つの尺度と離散・連続の区別**に応じた分析法
   - 今回扱う統計は，**要因に既に心当たりがある場合**の統計
   - 説明変数の尺度はそこまで多様ではない？
   - **コーディング**（変数にどのような数字を当てるか）という別の問題が関係



応答変数の分布：確率変数・確率分布の導入として

導入の導入：なぜ分散も大事か

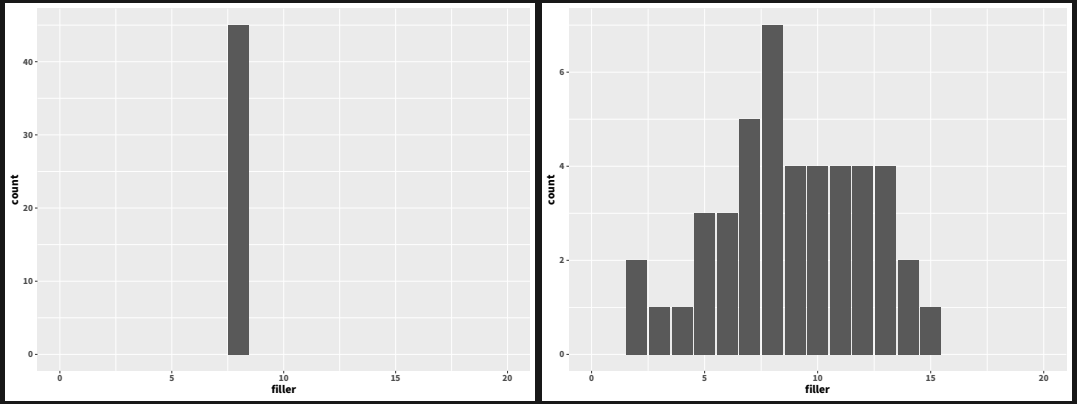
30 人に調査をして，ある時間内で何回「えーっと」（フィラー）を言うか数えた

- ある調査では，平均の「えーっと」の産出数が 8 回
- 別の調査では，平均の「えーっと」の産出数が 8.8 回

両者から得られたデータは同質？

導入の導入：なぜ分散も大事か

30 人に調査をして，ある時間内で何回「えーっと」（フィラー）を言うか数えた



(a) 平均 8 回，標準偏差 0 回

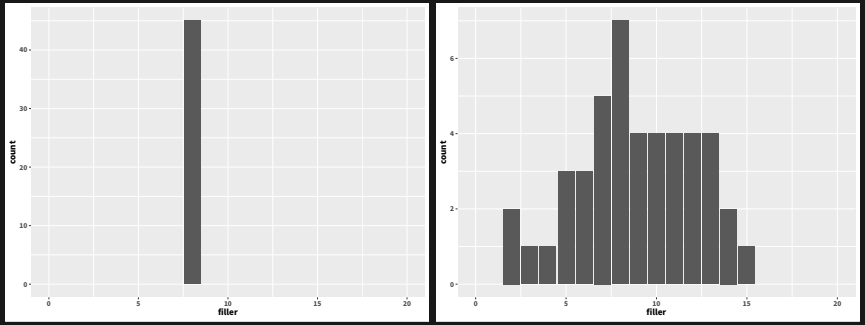
(b) 平均 8.8 回，標準偏差 3.3 回

Figure 3: フィラーの回数

ばらつきが大きく違う

導入の導入：なぜ分散も大事か

30 人に調査をして，ある時間内で何回「えーっと」（フィラー）を言うか数えた



(a) 平均 8 回，標準偏差 0 回　　**(b)** 平均 8.8 回，標準偏差 3.3 回

Figure 4: フィラーの回数

```
mean(m8_uni$filler); sd(m8_uni$filler)
```

```
## [1] 8
```

```
## [1] 0
```

```
mean(m8_pois$filler); sd(m8_pois$filler)
```

```
## [1] 8.844444
```

```
## [1] 3.254057
```

ああ

30 人に調査をして，ある文を何ミリ秒で読むか計測した

導入の導入：なぜ分散も大事か

- データが複数あり，どのデータの平均も等しかったとしても，データによってばらつきが異なる場合がある
  - というより，ほぼ毎回ばらつきは異なる
  - ばらつきの指標 = 分散

導入の導入：なぜ分散も大事か

- データが取りやすい値は異なる
  - 「えーっと」の例でも，「えーっと」を 3 回言う頻度よりも，10 回言う頻度の方が高い（標本 = 実測値の場合）
  - 「えーっと」を 3 回言う確率より，10 回言う確率の方が高い（母集団の場合）
  - **確率変数**：値それぞれに「その値の取りやすさ」（確率）が考えられるような項目とその具体的な実現形
    - 例：「えーっと」を何回言うか（「3 回言う」「10 回言う」）
    - 「事象」と「その事象の実現形」として捉えた方が理解しやすいかも…？
  - **確率分布**：個々の実現形としての確率変数に，どのくらいの確率が対応しているか
    - 変数（ここでは応答変数）の現れ方のパターン
    - 例：「えーっと」を 0 回言う確率が p_0，1 回言う確率が p_1，…10 回言う確率が p_{10}…
    - =「えーっと」を 0 回言うことに対して確率 p_0，1 回言うことに対して確率 p_1，…10 回言うことに対して確率 p_{10}…というように，値と確率が対応している
    - 「能動文・受動文の例」の場合，確率分布（確率変数と対応する確率の関係）は，どう記述できる？
    - 「読み時間」の場合，確率分布は，どう記述できる？

確率分布

尺度や型に応じて分布の違いがある

ここで，尺度や型を学んだ意味が出る

応答変数の尺度	応答変数の型	応答変数の取りうる範囲	確率分布	分散の平均との関
量的変数	離散型	0 以上，無限（$[0, \infty]$）	ポアソン分布	平均に近似
名義尺度	離散型	0 以上，有限	二項分布	平均の関数
量的変数	連続型	$[-\infty, \infty]$	正規分布	平均と無関係

標本集団の平均と分散

母集団の平均と分散をどう推定するか？

母集団の平均と分散を定義する

母平均 = 期待値

母分散

離散型と連続型

日本統計学会 (Ed.). (2017). 改訂版日本統計学会公式認定統計検定 2 級対応統計学基礎. 東京図書.