

# 消費者物価指数(CPI)への ウェブスクレイピングの活用について

令和元年11月13日

総務省統計局

物価統計室

## 1. ウェブスクレイピングの取組の概要

- ✓ 背景・経緯
- ✓ 取組の概要 など



## 2. 試行を通じて直面した課題と対応

- ✓ データ特性・技術的な課題
- ✓ データホルダーとの連携
- ✓ 実運用に向けた体制整備 など



## 3. まとめ



# 1. 取組の概要 | CPIとは

- CPI（消費者物価指数）は、全国の世帯が購入する様々な商品（財及びサービス）の価格変化を総合した指標
- 世帯の購入行動を代表する品目（計585）について、具体の売れ筋商品（銘柄）を選定し、全国のスーパーなど（約28,000事業所）の**実地調査**、各社の**ウェブサイト**、**POSデータ**により価格情報を収集
- 物価変動をとらえるためには、**同品質の商品の価格を追跡**※して把握することが重要

## 価格収集



統計調査員等

### スーパー、専門店などへの実地調査【約490品目】

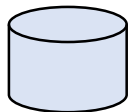
- ▶ 食品、飲料、生活家電、衣料品、医薬品、交通、教育、教養娯楽サービスなどの幅広い品目



総務省職員

### 各社ウェブサイト等から【約50品目】

- ▶ 店頭販売価格：コーヒー飲料（セルフ式）、ドーナツ（外食）、リサイクル料金、乗用車、ロードサービス料、有料道路料、鉄道運賃 など
- ▶ ネット販売価格：航空運賃、サッカー観覧料、テーマパーク入場料、ウェブコンテンツ利用料（電子書籍、動画配信、オンラインゲームなど含む）、振込手数料（ネットバンキング含む）、週刊誌、サプリメント（健康保持用摂取品）など



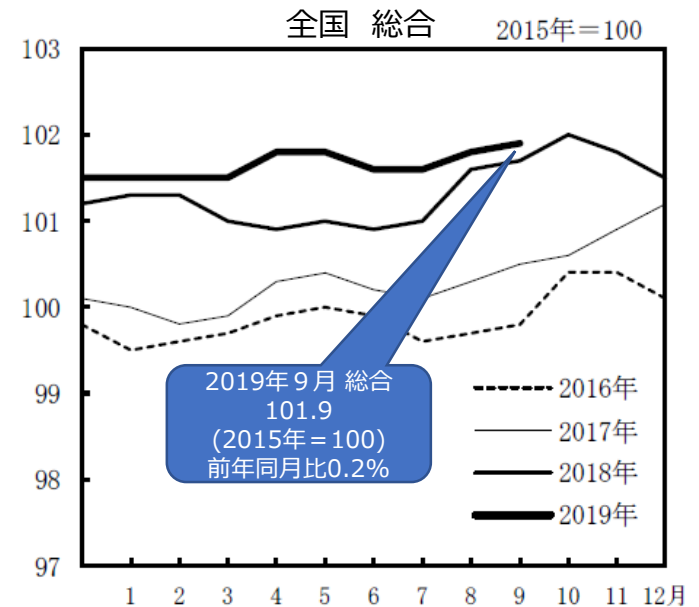
### POSデータ

- ▶ パソコン（デスクトップ型、ノート型）、カメラ

統計処理

## CPI

（毎月公表）













※ 売れ筋商品の変化に対応して調査する商品（銘柄）を変更する場合、品質の違いにより生じる価格差を除外するため、商品の機能、規格、容量等の特性情報を用いて品質調整を行っている。【参考1】

# 1. 取組の概要 | ネット販売の進展

- 現在、ネット購入割合が高い品目や、ネット上で対面販売価格と同一価格が把握できる品目を中心に、ネット販売価格をCPIに反映【参考2】
- 近年のネット販売の増加や、ネット情報の収集技術の進展を背景に、ウェブスクレイピングの有用性の拡大が期待

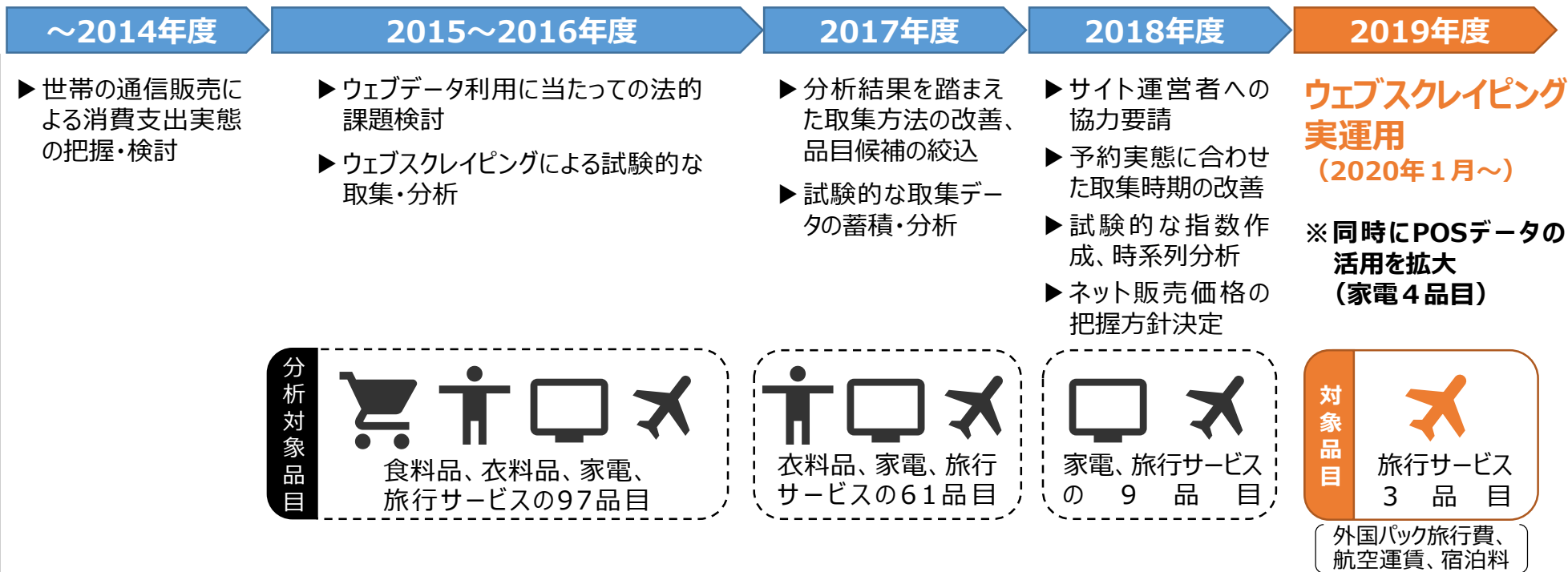
## 主な商品・サービスのネット販売の状況

	EC市場規模 注1 (2018年)	ネット購入割合 注2 (2014年)	価格情報の入手 が可能な情報源
食料 	16,919億円 (食品、飲料、酒類)	0.6%	 POS (ネット販売を含まない)  Web
被服及び履物 	17,728億円 (衣類・服飾雑貨等)	3.1%	 Web
教養娯楽用 耐久財 	16,467億円 (生活家電、AV機器、PC・周辺機器等)	10.5%	 POS  Web
旅行サービス 	37,186億円	40.3% (航空運賃)	 Web

※ 「EC市場規模」と「ネット購入割合」は、調査方法が異なる結果を引用していることから、単純な数値の比較はできないことに注意を要する。

注1 経済産業省「平成30年電子商取引に関する市場調査」【参考3】

注2 総務省「平成26年全国消費実態調査結果」（二人以上世帯）における「通信販売(インターネット)を利用した購入割合」【参考4】



**統計委  
答申**  
(2012年1月)

☞ 全国物価統計調査の中止に伴い、同調査で把握していた**通信販売価格**（インターネット、カタログ販売等）について、その把握の必要性及び技術可能性について検討

**経済財政  
諮問会議**  
(2015年10月)

☞ ネット通販市場が拡大しているにもかかわらず、家電をはじめほとんどの商品について**ネット販売価格が加味されていない**

**統計改革  
の基本方針**  
(2016年12月)

☞ **インターネット販売価格の更なる捕捉及び**（中略）**採用の可否を検討**

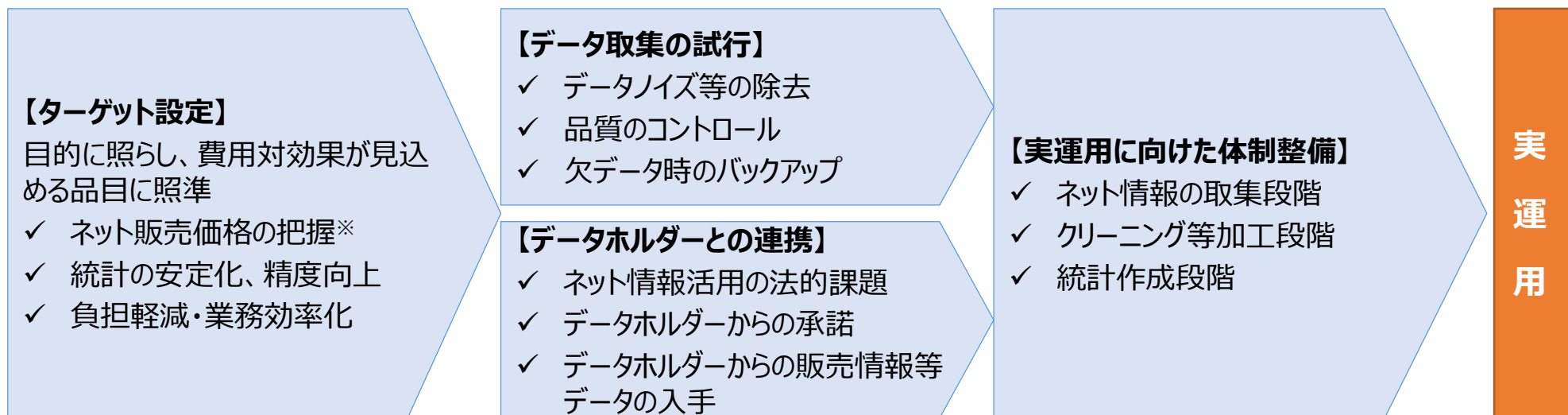
**公的統計  
基本計画**  
(2018年3月)

☞ **インターネット販売価格の採用の可否を検討**

## 2. 試行を通じて直面した課題と対応

- ウェブスクレイピングにより、①**ネット販売価格の直接把握**、②**膨大な量の価格データによる統計精度の向上**、③**報告者負担軽減・業務効率化** が期待でき、現行調査を補完できる可能性
- CPIでは、品目ごとにネット販売の状況や活用できるネット情報などが異なるため、試行を重ねながら費用対効果を見極めていく必要

### ウェブスクレイピング実運用に向けた大まかな検討プロセス



※ 一般にウェブスクレイピング情報は、店頭・ネット販売価格全体を代表していないことにも留意

## 2. 課題と対応 | データ収集の試行 (外国パック旅行費)

- 大手旅行会社が提供するネット販売価格を捉えることで、「外国パック旅行費」全体の価格動向を適切に把握可能
- 旅行会社から外国パック旅行の販売実態に関する詳細な情報提供が得られたことにより、膨大な数の商品情報からノイズを除去し、物価動向把握で前提とする同品質の商品の価格追跡が可能
- 従前に比べ、カバレッジ拡大（旅行会社、価格収集期間、プラン数など）が可能となり、統計精度の向上に寄与

### 旅行会社からのヒアリング結果

- ✓ 「外国パック旅行」（消費者向け）の販売数量は、代理店販売がネット販売を上回っているものの、そのプランのほとんど（富裕層向けプランなどを除く）がネットに掲載
- ✓ ネット販売は価格比較サイトなどでも行われているが、大半が旅行会社経由（比較サイトで購入する場合も旅行会社のサイトに移行して購入される）

### ウェブスクレイピングの試行

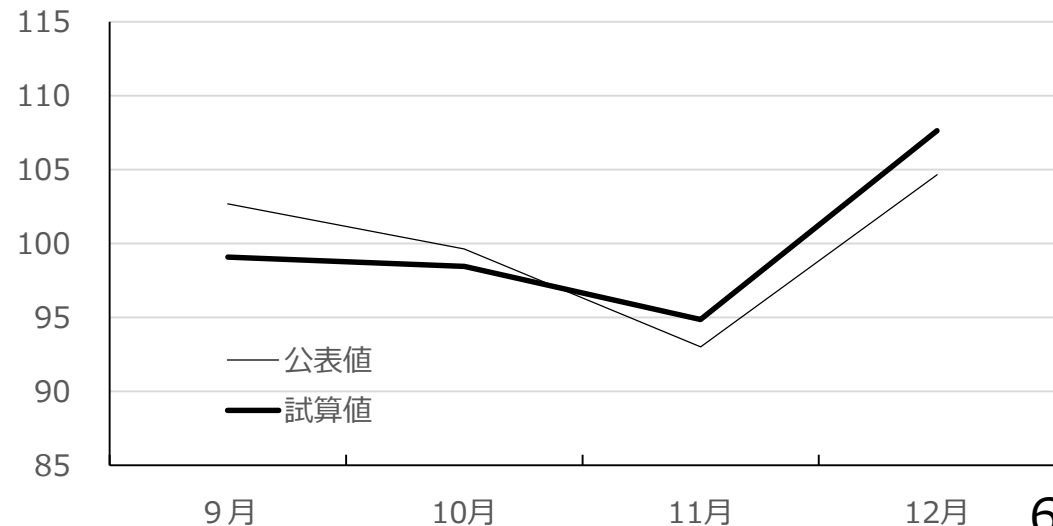
- 期間：2018年9～12月出発分  
（予約が多い2～3か月前に、出発月の全日の価格を取集）
  - 旅行先：日本人訪問者数の多い10の国・地域
  - 旅行会社：大手旅行業者複数社  
（主要旅行業者合計取扱額シェア50%以上）
  - 航空会社：安定してプラン提供される複数社
  - プラン：安定提供されるグレードの宿泊施設、フリープラン
- ※ ノイズ除去のイメージは次ページ参照

### 分析に用いた価格数

	9月出発分	10月分	11月分	12月分
取集価格数	245,288	181,533	191,470	203,142
試算に用いた価格数	127,747	101,211	108,568	107,402

指数（期間平均 = 100）

### 現行CPIと試算値の比較



## 2. 課題と対応 | データ収集の試行（外国パック旅行費）（続き）

### データクリーニングのイメージ

○○旅行.com

トップ > 海外ツアー > ハワイ > 成田発 ホノルル

検索

成田発 ホノルル <ホテルは決定済み！専用バス・トロリーに滞在中乗り放題！到着日空港～ホテル間の送迎付き！> 直行便、フリープラン、バス乗り放題、アンバサダーホテルワイキキ（部屋指定なし）、6日間、燃油サーチャージ込み

2名1室 1

価格：¥209,800（税込）

購入する

### HTMLソース

価格

```
<th><img id="ctl00_SearchSiteHolder_Ext" alt="日程（6日間）" data-bbox="350 421 450 445" style="border: 1px solid black; width: 100px; height: 20px; vertical-align: middle;"/><td class="travel_price_contents"><p><span class="price">&yen;209,800</span></p><p><span class="note2">※空港諸税等は含まれておりません。</span></p><th><td class="travel_surcharge_contents"><p><span class="surcharge">旅行代金に含む</span></p><p><span class="note2">※船舶燃油特別付加運賃は含まれておりません</span></p><span class="name"><span class="em01">アンバサダー ホテル ワイキキ(部屋指定なし)/ホノルル (ハワイ) </span></span></td></tr>
```

宿泊ホテル名

### データセット

(2019年9月収集データのクリーニング例)

217万価格

- ① ウェブデータ収集（条件検索）
  - ・ 出発地・目的地
  - ・ 日程：○日間
  - ・ プラン：フリープラン
  - ・ 2名1室利用、大人1名、
  - ・ 全航空会社、全宿泊施設グレード
  - ・ 収集月の1～8か月後の月の全日価格

158万価格

- ② 以下のデータの削除
  - ・ 料金「-」
  - ・ プラン名又はホテル名に複数都市名（100パターン）
  - ・ ビジネス、プレミアムエコノミー、指定以外のプラン等
- ③ 航空会社名の統一（ANA⇒全日空、JAL⇒日本航空 など）

117万価格

- ④ ホテルグレード格付け及び旅行会社間のグレード統一
    - ・ SL>L>A>B>C>D>E>Fの8グレードを約2,000施設※に対して格付け
  - ⑤ 利用が低いグレードの削除
    - ・ SL、D、E、F（4グレード）
- ※ 同一ホテルでも月によって表記が異なることがある  
(SSホテル、エスエスホテル、エスーエス など)

22万価格

- ⑥ 売れ筋の「旅行先、航空会社、ホテルグレード、予約時期（1～8か月前）」の組合せのデータとそれ以外を分離し、売れ筋のみを利用



## 2. 課題と対応 | データ収集の試行（航空運賃※） ※ 国内

- 大手航空会社のサイトによるネット販売価格を捉えることで、国内航空券全体の価格動向を適切に把握可能
- 航空会社から航空券の販売実態に関する詳細な情報提供が得られたことで、膨大な数の価格情報からノイズを除去し、同品質の商品の価格追跡が可能
- 従前に比べ、カバレッジ拡大（路線数、価格収集期間）が可能となり、統計精度の向上に寄与

### 航空会社からのヒアリング結果

- ✓ 航空券（消費者向け）の販売数量は、ネット販売が最も多い
- ✓ 旅行会社販売、電話販売の航空券は、ネット販売価格と同一（各システム間で連携されている）

### 分析に用いた価格数

	1月搭乗分	3月分	6月分	9月分
収集価格数	575,530	543,332	670,646	978,096
試算に用いた価格数	160,614	159,003	197,451	212,697

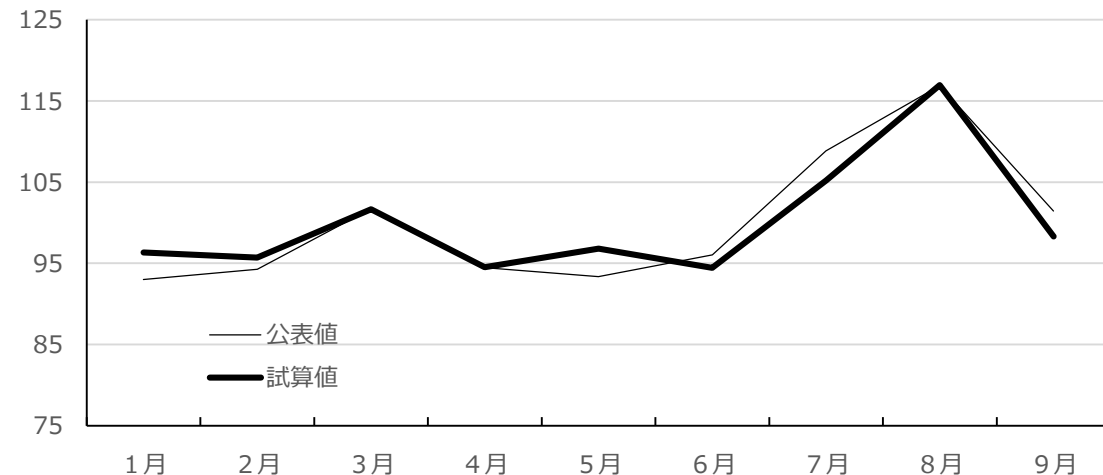
※ 紙面の都合から一部を抜粋。9月は15路線の価格を収集

### ウェブスクレイピングの試行

- 期間：2018年1～9月搭乗分（割引運賃の種類に合わせ、搭乗する75日前、55日前、45日前及び28日前における価格を収集）
  - ※ 価格収集タイミングについては次ページ参照
- 運賃種別：普通・往復、各種割引
- 路線：旅客数の多い10路線の全便
- 航空会社：大手航空会社複数社（旅客数シェア90%以上）

指数（期間平均＝100）

### 現行CPIと試算値の比較



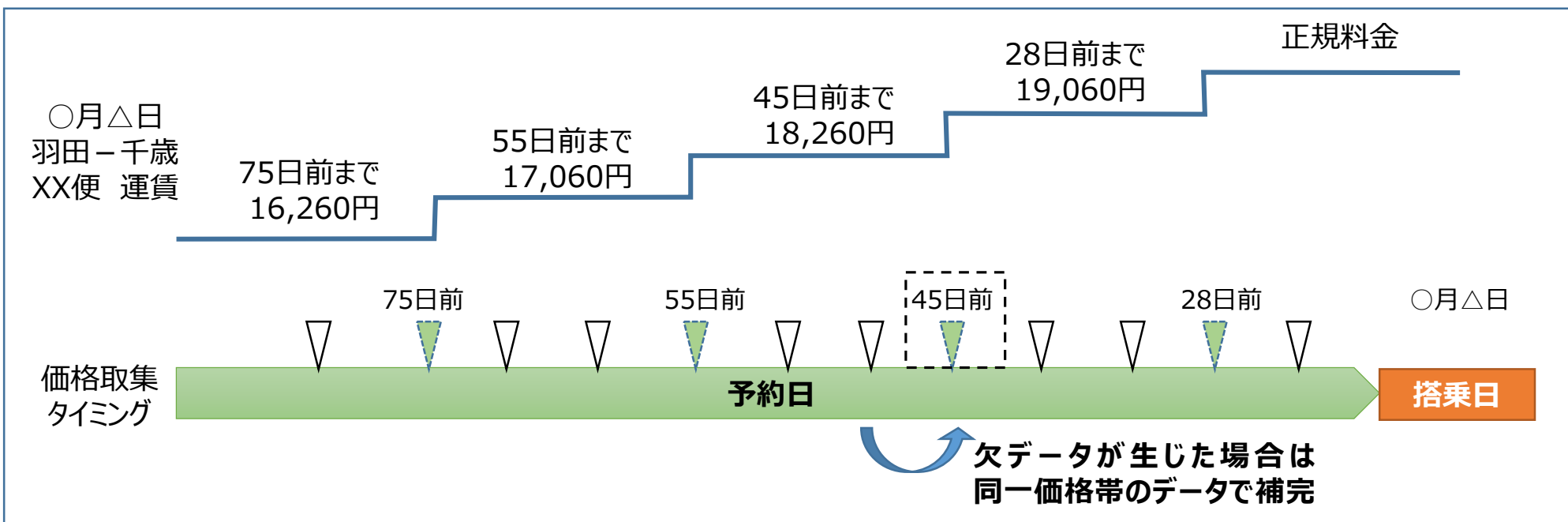
## 2. 課題と対応 | データ収集の試行（航空運賃）（続き）

- 航空運賃の価格体系を踏まえ、効率的な価格収集日を設定
- 定期・不定期のサイト変更、アクセス集中、システム障害などに伴い、指定日の価格データが収集できない場合に備え、同一価格帯の価格データを複数収集して、欠データを自動補完することで対応

### 【航空運賃の基本的な価格体系】

- ✓ チケットの予約日から搭乗日までの期間が長いほど、料金は安い
- ✓ 同一割引価格帯（75日前など）であれば、区間・便ごとに料金設定され、販売状況によって価格は変動
- ✓ 届出された料金間を移動し、同一割引価格帯であれば変動はそれほど大きくない（超繁忙期除く）

### 価格収集のイメージ



## 2. 課題と対応 | データ収集の試行（宿泊料※） ※国内

- 旅行予約サイトによるネット販売価格を捉えることで、宿泊料全体の価格動向を適切に把握可能
- 膨大な数の価格情報からノイズを除去し、同品質の商品の価格追跡が可能
- 新たな回答負担を課すことなく、カバレッジ拡大※（毎日の宿泊価格の把握等）が可能となり、統計精度の向上に寄与

※ これまでの宿泊料調査は、全国320施設を対象として、毎月5日を含む週の金・土の宿泊価格を調査

※ 対象施設は、これまでの320から400へ拡大予定（リサンプリングによる標準誤差率の分析結果）

### ウェブスクレイピングの試行

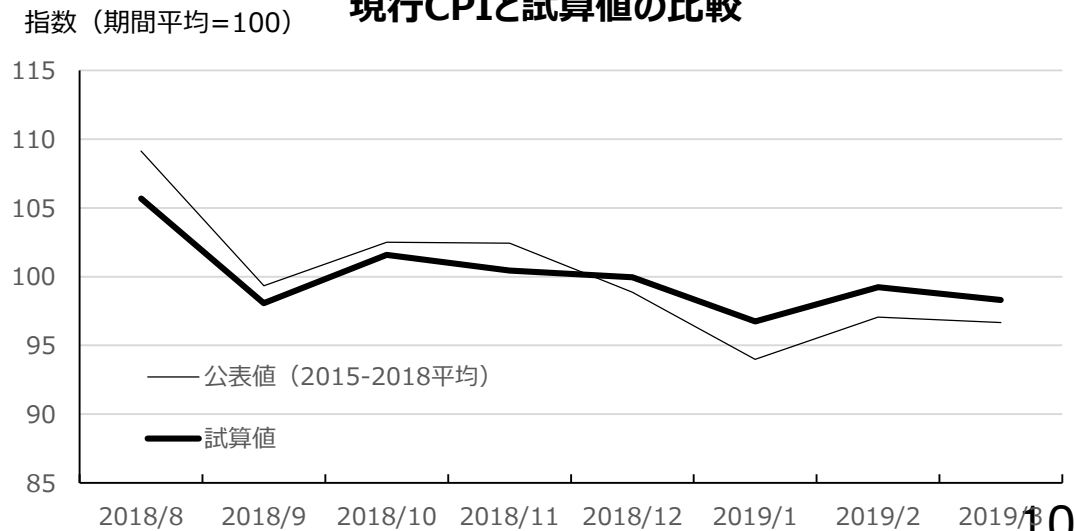
- 期間：2018年7月～2019年3月宿泊分（宿泊日の1～2か月前に価格を収集）
- 価格収集サイト：大手旅行業者複数社による旅行予約サイト  
（主要旅行業者合計取扱額シェア50%以上）
- 宿泊施設：現行の宿泊料調査の対象（320施設）を基本として、観光庁の統計などを参考に施設を選定
- プラン：「和室・1泊2食付」、「洋室・1泊朝食付」  
※ 出張向けのビジネスプラン、高価格帯のスイートルームなど、CPIから除外すべき価格は、外れ値処理  
※ 満室等のため価格表示されなかった場合は回帰補完

### 分析に用いた価格数

	7月宿泊分	9月分	12月分	3月分
収集価格数	519,181	1,269,219	1,887,649	1,290,077
試算に用いた価格数	516,380	1,263,509	1,878,600	1,284,253

※ 紙面の都合から一部を抜粋

### 現行CPIと試算値の比較



## 2. 課題と対応 | データ収集の試行（衣料品）

- 価格データが掲載されているサイトには、販売店舗を持つ個別企業のサイトや、多数の出展者の商品を扱うEC事業者のサイトが存在
- EC事業者のサイトでは、幅広い商品の情報が掲載されているものの、商品説明（商品名、素材など）や表記方法などの質・量が出展者によってバラバラな傾向

⇒ 同品質の商品の価格追跡が必要なCPIにおいては、ウェブスクレイピング活用は難易度が高い

商品名称にPR  
情報等が混在

特定の商品名・  
コードが存在しない

商品の品質に関する  
情報の表記が不統一  
で情報量が膨大

### 衣料品ECサイトの掲載情報の例（背広服）

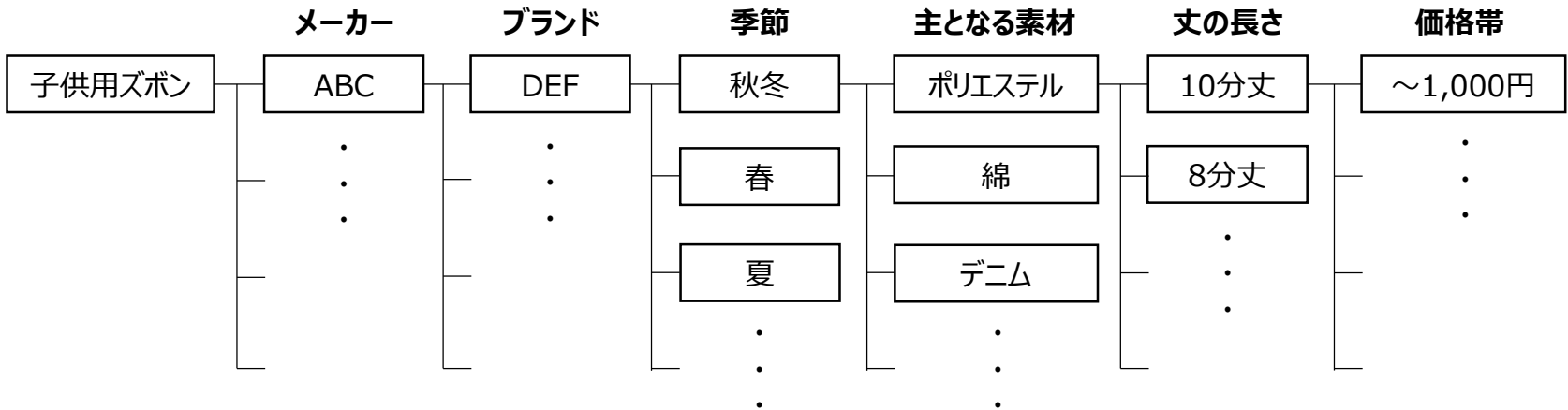
商品名	商品コード	出展者	商品説明の一部	商品説明の文字数
2つボタンスーツ ピンストライプ	—	A	…（略）…【生地】 Biella Finish イタリア技術者の指導を受け、原料選定から紡績、織布、仕上げまでの加工に、イタリアのノウハウを導入して作り上げた生地。高度な技術を具現化することで、イタリア製生地に遜色のない上質な生地に仕上がっています…（略）…【仕様：ジャケット】2つボタン／本切羽／センターベント／胸ポケット／腰ポケット×2／総裏仕立て／内ポケット×2【仕様：パンツ】ノータック／スリムフィットテーパード／エキストラローライズ／サイドポケット×2／ヒップポケット×2／ウォッチポケット／前ヒザのみ裏地あり【季節】秋冬モデル【洗濯表示】ドライオンリー ● 体型 YA（Drop8：スリム） A（Drop6：普通） AB（Drop4：ややがっすり） ● 身長【3】160cm …（略）…	約1,500文字
1着は持っていたい!ウエストのアジャスター付きがうれしいシングルフォーマルスーツ メンズ	—	B	ビジネスばかりでなく、礼服・フォーマルにも使える黒無地スーツはマストアイテム。ウエストにはアジャスターを装備し、左右で±3cmずつ調整できるので、長時間の座り姿勢のときや、食後にお腹が張っているときなども快適。ベーシックなスーツなので、年齢も関係なく着用できるのがうれしい。■表地:毛50%ポリエステル50%(織物・背抜き)裏地:ポリエステル100% ▼スラックスの裾は半仕上げになっています。ご家庭でお好みの寸法にお直しく下さい。ご家庭で簡単にできる裾上げテープを販売しておりますのでご利用ください。●ウエストアジャスター機能付き ●ドライのみ(ジャケット・パンツ) ●ミャンマー製…（略）…	約300文字
秋冬物ウォッシュブル2ボタンスーツ【ブラック / ストライプ織柄】	—	C	…（略）…ご家庭でお洗濯可能! 2ボタンスーツ (8250-03) Item Information Dady Costa 2ボタンスーツのご紹介です。ウォッシュブル生地を採用し、ご家庭でお洗濯が可能なスーツです。スタイルは極端にラインを細く強調するのではなく、着心地を重視した2ボタンスーツで、細身のスーツが苦手なお客様にもご満足いただけるスーツになっています。ブラック地のストライプ織柄でオーソドックスながら品の良さがあり、どのような場面でも対応可能で幅広く着用可能なスーツです。ズボンの折り返し目が見えない耐久折り返し加工や、貴重品を守るフラップ付き内ポケットなど、充実した仕様となっております。お手頃な価格ですので、スーツを何着も必要なお客様や家計のやりくりで苦労している奥様にも喜んでいただける商品です。Dady Costa Washable 2 Button Suit 8250-03 WEB Price (税抜き) 19,000 Information Size List A4 A5 A6 A7 AB4 AB5 AB6 AB7 BB4 BB5 BB6 BB7 Cloth Information ブラック / ストライプ織柄 ウール55% ポリエステル45% 日本製生地 ブラックのバイアス織柄地に4mm間隔でドライブが入っています…（略）…	約2,000文字

商品の同一性・同質性の判別が困難

## 2. 課題と対応 | データ収集の試行 (衣料品) (続き)

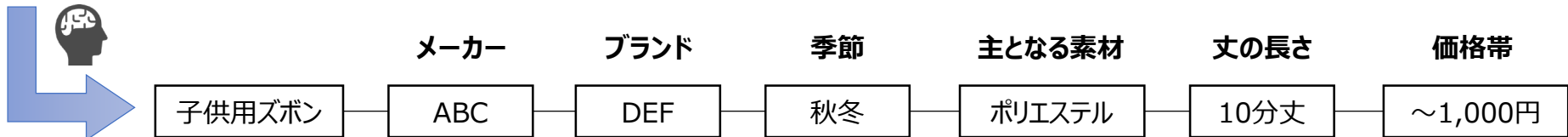
- 衣料品ネット販売データの製品区分などの格付けのため、AI (機械学習) の活用可能性について検証中 (2019年度～)

### 分類イメージ



### AI機械学習 イメージ

商品名 : **ABC 長ズボン** キッズ 子供服 男の子 **2～5歳 薄手のデニム** ジーンズ **青**  
 説明 : 薄手なので**季節の変わり目**にぴったり! キッズ 薄手のデニムジーンズ 長ズボン(2～5歳 : **90・95・100・110**)  
 薄手のデニムが入荷しました。薄手なので、季節の変わり目や蚊の多い公園など**アクティブ**に過ごすときに最適!  
 半ズボンから長ズボンに変えるのを嫌がるお坊ちゃまにも。  
**【キッズ】【男の子】素材 ポリエステル100% サイズ 2T,3T,4T,5T**  
 注意点 ○洗濯機、乾燥機は弱で。○漂白剤不可。○**色落ち**する場合がございますので、最初の洗濯時は白いものとは分けてください。



## 2. 課題と対応 | データホルダーとの連携

- ネット情報の利用に当たっては、法的な課題・留意点を踏まえた上での対応が必要  
(ウェブスクレイピングを禁止しているサイトもあり)
- 商品情報の提供のみならず、ウェブスクレイピングに関して、データホルダーの理解・協力が不可欠
  - ✓ ネット情報の収集及び統計作成に関して、**全対象企業から承諾**
  - ✓ ネット情報の収集に当たっては、サーバーへの負荷を考慮して、企業側から提示のあった**アクセス時間帯・頻度（アクセス回数/秒）の制限、IPアドレスの事前登録**などの対応を実施

※ 諸外国においても、ウェブスクレイピングに当たって事前承諾を得ているとする国が多い。【参考5】

主な留意点	法律事務所等において示されている考え方 <sup>注)</sup>
(a) ネット情報の利用は、著作権に関し問題になる可能性はないか	<p>「価格」「商品名」「配送料」「販売店名」「発送企業名」「出品者情報」等の収集するウェブ情報は、<b>著作権法上保護される「著作物」には該当しない</b></p> <p>※ 「著作物」とは、「思想又は感情を創造的に表現したものであって、文学、学術、美術又は音楽の範囲に属するものをいう。」（著作権法第2条第1項第1号）</p>
(b) サイトの利用規約等により、法的拘束力が発生する可能性はないか	<p>利用者に制限をかけずにサービス提供を行っているサイトの場合、当事者間での意思の合致があったとはいえ、<b>合意が成立しているとは言い難く拘束力が働くとまではいえない</b></p> <p>※ 会員登録等のように、利用規約等（定型約款）への合意を条件としてサービスの提供を行っているサイトの場合、合意したものとみなされ、法的な拘束力が有効となる（民法上の債務不履行・不法行為が生ずる可能性）</p>
(c) アクセスの程度によっては、業務妨害となる可能性はないか	<p>多かれ少なかれサイト側のサーバに過度の負荷をかけ、程度を超えてアクセスすると、場合によっては、<b>アクセス先のシステムにも何らかの支障を生じるおそれ</b>（刑法上の偽計業務妨害罪に該当する可能性）</p>

注) (a)(b)は1法律事務所に個別に相談。(c)は統計局職員が複数の法律事務所HPを確認。

## 2. 課題と対応 | 実運用の体制

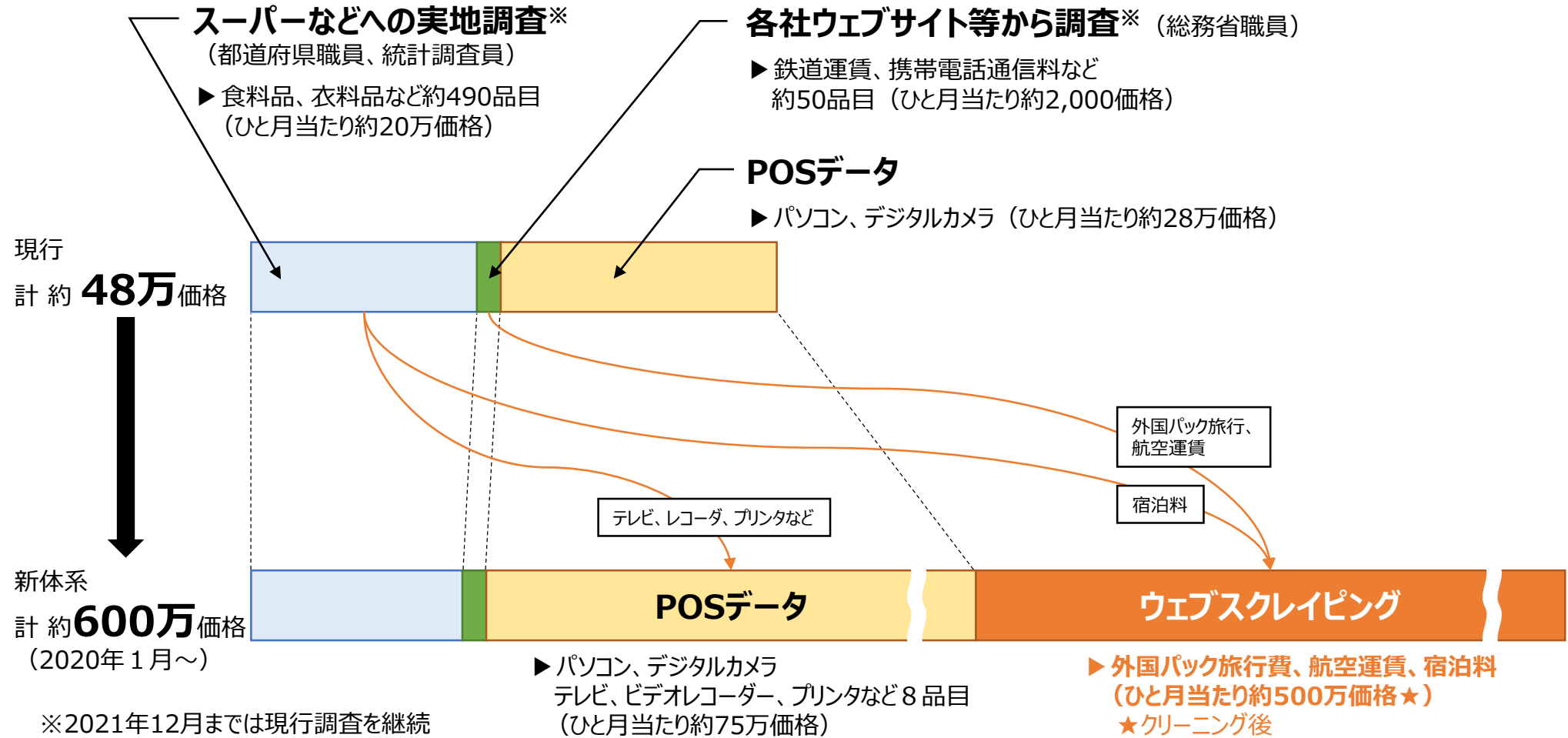
- CPIは月次統計であり、ウェブスクレイピングによる膨大なデータの収集、データノイズ等の除去、統計作成に至るまでの一連の作業を短期間で完了する必要
- 業務の性質、ハードウェア・ソフトウェアの環境、人的リソースなどを考慮して、専門事業者と連携して業務運営に当たることが必要

### 統計作成までの業務のリソース配分

主な作業工程	リソース配分に当たり主に考慮した点	業務実施者
① ネット情報の収集	<ul style="list-style-type: none"><li>● 複数のウェブスクレイピング専門業者の存在</li><li>● Webプログラム言語の精通の必要性</li><li>● 原則、収集条件設定は変わらない (※ サイト統合などの場合に変更が発生)</li></ul>	民間リソースを活用
② データクリーニング	<ul style="list-style-type: none"><li>● 膨大なデータを短期間でクリーニング処理可能なハードウェアが必要</li><li>● 処理プログラム言語の精通の必要性 (※ 職員は既に処理プログラム言語にある程度精通)</li><li>● データ表記方法の変更、ホテル取扱いの変更などの場合、即時の柔軟な対応が必要</li></ul>	変更のない定型処理のみ 民間リソースを活用
③ 統計作成	<ul style="list-style-type: none"><li>● サイト運営者から個別に受理した販売シェア等の情報の保護</li><li>● 処理プログラム言語の精通の必要性 (※ 職員が既に処理プログラム言語にある程度精通)</li><li>● 欠損時の処理方法・決定及び即時の実行</li></ul>	総務省

### 3. まとめ | 新たなフェーズへ

- 2020年1月から、ウェブスクレイピングによる価格収集（及びPOSデータの活用拡大）を開始することにより、CPIの情報源が膨大なデータ量となり、新たなフェーズに突入
- 2021年度から、これらのデータを取り込んだCPI（2020年基準）を公表開始





### 3. まとめ | ウェブスクレイピングの特徴

- ウェブスクレイピングにより、①**ネット販売価格の把握拡大**、②**膨大な量の価格データによる統計精度の向上**、③**報告者負担軽減・業務効率化** が期待
- ウェブスクレイピングの活用拡大に当たっては、品目ごとのネット販売の進展状況や、ネット情報のデータ特性などに照らし、費用対効果を慎重に見極めていくことが必要

	実地調査の特徴	ウェブスクレイピングの特徴
把握できる販売形態	<ul style="list-style-type: none"> <li>● 店頭での実際の販売価格を把握可能（ネット販売価格は得られない）</li> <li>⇒ 店頭販売が主流の商品については正確な実態把握が可能</li> </ul>	<ul style="list-style-type: none"> <li>● ネット販売価格を把握可能（店頭販売価格と異なる可能性）</li> <li>⇒ ネット販売が主流の商品については正確な実態把握が可能</li> </ul>
データ数	<ul style="list-style-type: none"> <li>● 得られるデータ数が限定的</li> <li>● 調査日が限定的（カレンダー要因が発生する場合あり）</li> <li>⇒ 統計に振れが生じる場合がある</li> </ul>	<ul style="list-style-type: none"> <li>● 膨大な量の価格データを収集可能</li> <li>● 日次価格の収集が可能</li> <li>⇒ より統計が安定することが期待できる</li> </ul>
報告者負担、業務負担	<ul style="list-style-type: none"> <li>● 報告者負担が大</li> <li>● 職員や調査員の業務量大</li> <li>● 人件費などの費用が大</li> </ul>	<ul style="list-style-type: none"> <li>● 報告者負担はほぼゼロ</li> <li>● 収集自動化により業務の効率化が可能</li> <li>● 比較的安価</li> </ul>
その他の課題		<ul style="list-style-type: none"> <li>● バイアスが存在（店頭・ネット販売全体を代表していない）</li> <li>● 法外な価格や不要な情報などのデータノイズが混在</li> <li>● サイト更新やシステム障害などによりデータ収集できない場合の備えが必要 など</li> </ul>

### 3. まとめ | 今後の取組拡大の方向性

- ネット販売の増加や、ネット情報の充実・その収集技術の進展を背景に、ICTを活用した価格収集の有用性が拡大していくことが見込まれる
- デジタル経済が進展する中で、ウェブスクレイピングやPOSデータ等のビッグデータの活用、AIの活用、その他のICTによる作業自動化などの導入により、統計精度向上と報告者負担軽減・業務効率化の両立を引き続き追求

#### 中長期的な検討対象

	品目例		候補となる価格収集方法
POSデータが入手可能と考えられる品目	家電	電気炊飯器、ルームエアコン など	POSデータ
ネットスーパーのウェブサイトからの価格収集が可能と考えられる品目	食料品、医薬品、日用品	うるち米、感冒薬、殺虫剤 など	ウェブスクレイピング
<ul style="list-style-type: none"> <li>・ 価格変動が少なく、</li> <li>・ ウェブクローラによる定期的アクセス・差分検知により、効率的に価格変動を把握可能と考えられる品目</li> </ul>	《調査員調査品目》 コーヒー(外食(セルフ))、焼肉(外食) など 《都道府県調査品目》 行政証明書手数料、パスポート取得料など		ウェブサイトからの価格収集 (調査員/都道府県調査品目から総務省調査品目への移行を検討)

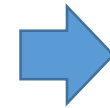
以下參考資料

# 【参考1】品質調整の例

- 物価変動をとらえるためには、**同品質の商品の価格を追跡**して把握することが原則
- 売れ筋商品の変化に対応して調査する商品（銘柄）を変更する場合、品質の違いにより生じる価格差を除外するため、商品の機能、規格、容量等の特性情報を用いて品質調整を実施

例1) 容量比による換算 … 同一商品で容量だけ減った場合

	前月	当月
商品A	1000g 215円	-
商品A'	-	900g 207円



	前月	当月
商品A	1000g 215円	1000g 230円
		[207円×1000g/900g]

例2) 回帰式による換算 … 内蔵メモリのみが異なる後継機（廉価版）が販売されその商品が売れ筋となった場合

	前月	当月
商品A	64GB 2万円	-
商品A'	-	32GB 1.2万円
商品A''	-	128GB 2.4万円

※商品A、A'、A''の品質の違いは内蔵メモリのみ



[回帰式による推計]

$$\begin{cases} 1.2万 = 32a + \beta \\ 2.4万 = 128a + \beta \end{cases} \quad \therefore a=125, \beta=0.8万$$

$$y = 125x + 0.8万$$

	前月	当月
商品A	64GB 2万円	64GB 1.6万円
		[125円×64GB+0.8万円]

例3) ヘドニック法による換算 … 品質向上が著しく製品サイクルが短いパソコンなど

	前月	当月
商品A	P万円 HDD、メモリ、ソフト…	-
商品B	-	P'万円 HDD、メモリ、ソフト…



[重回帰式による推計]  $\ln p_T = \alpha_t + \beta_t \delta_{T,t} + \sum_k \gamma_{t,k} x_k$

$p_T$ : 販売価格  $\delta_{T,t}$ : 販売時点ダミー (前月=0, 当月=1)

$x_k$ : 特性量 (HDD記憶容量、メモリ容量、バンドルソフトの有無など)

品質調整済前月比 = $\exp[\hat{\beta} - 1] \times 100(\%)$
---

## 【参考2】 現行のCPIにおけるネット販売価格の採用品目

- 現在、ネット購入割合が高い品目や、ネット上で対面販売価格と同一価格が把握できる品目は、ネット販売価格をCPIに反映

- **ネットショッピングによる購入割合が高い品目**（ネット販売価格を採用）

「健康保持用摂取品 A（※1）」、「健康保持用摂取品 B（※2）」、「航空運賃」

※1 マルチビタミン ※2 青汁

- **POSデータを利用している品目**

対面販売価格及びネット販売価格の両方が含まれているPOSデータを利用して製品間の価格差のうち品質に起因する部分を調整（ヘドニック法）した指数を作成

「パソコン（デスクトップ型）」、「パソコン（ノート型）」、「カメラ」

- **インターネットを通じて価格を調査（対面販売価格と同価格）している品目**

「振込手数料」、「レンタカー料金」、「携帯電話機」、「通信料（携帯電話）」、「コンパクトディスク」、「月刊誌」、「週刊誌」、「映画観覧料」、「演劇観覧料」、「サッカー観覧料」、「プロ野球観覧料」、「テーマパーク入場料」、「インターネット接続料」、「ウェブコンテンツ利用料」、「ハンドバック B」、「モップレンタル料」、「化粧クリーム A」、「化粧水 A」、「乳液 A」、「ファンデーション A」、「口紅 A」 など

# 【参考3】消費者向けネット販売の市場規模（平成30年）

サービス系分野		2013年 EC市場規模		2018年 EC市場規模		備考
①	旅行サービス	24,415億円		37,186億円		CPIに一部採用済（レンタカー料金など）、 <b>2020年から対象品目を拡大</b>
②	飲食サービス	935億円		6,375億円		<b>2020年から対象品目を拡大（宅配水）</b>
③	チケット販売	2,870億円		4,887億円		CPIに一部採用済（映画観覧料、テーマパーク入場料など）
④	金融サービス	6,447億円		6,025億円		CPIに一部採用済（振込手数料）
⑤	理美容サービス	※下記⑥金額に含む		4,928億円		
⑥	その他（医療、保険、住居関連、教育等）	6,043億円		7,070億円		CPIに一部採用済（通信料（携帯電話）、モップレンタル料など）

物販系分野		2013年 EC市場規模		2018年 EC市場規模		備考
		EC化率		EC化率		
①	食品、飲料、酒類	9,897億円	1.58%	16,919億円	2.64%	
②	生活家電、AV機器、PC・周辺機器等	11,887億円	22.67%	16,467億円	32.28%	CPIに一部採用済（パソコン、携帯電話機など）、 <b>2020年から対象品目を拡大</b>
③	書籍、映像・音楽ソフト	7,850億円	16.51%	12,070億円	30.80%	CPIに一部採用済（月刊誌・週刊誌、コンパクトディスクなど）
④	化粧品、医薬品	4,088億円	3.80%	6,136億円	5.80%	CPIに一部採用済（化粧品（化粧クリーム等）、健康保持用摂取品）
⑤	生活雑貨、家具、インテリア	9,638億円	13.17%	16,083億円	22.51%	
⑥	衣類・服装雑貨等	11,637億円	7.47%	17,728億円	12.96%	CPIに一部採用済（ハンドバッグ）
⑦	自動車、自動二輪車、パーツ等	1,675億円	1.87%	2,348億円	2.76%	
⑧	事務用品、文房具	1,354億円	23.30%	2,203億円	40.79%	

デジタル系分野		2013年 EC市場規模		2018年 EC市場規模		備考
①	電子出版（電子書籍・電子雑誌）	938億円		2,783億円		CPIに採用済（ウェブコンテンツ使用料（電子書籍、有料音楽配信、有料動画配信、オンラインゲームを含む））
②	有料音楽配信	416億円		645億円		
③	有料動画配信	597億円		1,477億円		
④	オンラインゲーム	8,423億円		14,494億円		
⑤	その他	646億円		984億円		

# 【参考4】世帯の消費支出におけるネット購入の割合（平成26年）

※中分類は通信販売割合が4%以上を表示

費目・大分類		消費支出	うちネット購入	ネット購入割合	備考
食料		71,914円	440円	0.6%	
住居		8,988円	22円	0.2%	
光熱・水道		2,602円	0円	0.0%	
家具・家事用品		10,396円	433円	4.2%	CPIに一部採用済（モップレンタル料）
うち	家庭用耐久財	3,037円	201円	6.6%	
	室内装備・装飾品	965円	49円	5.1%	
	寝具類	920円	50円	5.4%	
	家事雑貨	2,373円	97円	4.1%	
被服及び履物		14,147円	434円	3.1%	
うち	和服	173円	10円	5.8%	
保健医療		5,157円	200円	3.9%	
うち	健康保持用摂取品	875円	118円	13.5%	CPIに採用済
交通・通信		21,173円	281円	1.3%	CPIに一部採用済（携帯電話機、通信料（携帯電話）など）
うち	交通	4,628円	196円	4.2%	CPIに一部採用済（レンタカー料金など）
	<b>航空運賃</b>	<b>419円</b>	<b>152円</b>	<b>40.3%</b>	CPIに採用済、2020年から路線・便などのカバレッジを拡大
教育		2,936円	7円	0.2%	
うち	教科書・学習参考教材	70円	3円	4.3%	
教養娯楽		20,024円	809円	4.0%	CPIに一部採用済（映画観覧料、ウェブコンテンツ使用料など）
うち	<b>教養娯楽用耐久財</b>	<b>1,449円</b>	<b>152円</b>	<b>10.5%</b>	CPIに一部採用済（パソコン・カメラ）、2020年から対象品目を拡大
	教養娯楽用品	6,199円	312円	5.0%	CPIに一部採用済（コンパクトディスク）
その他の消費支出（諸雑費）		13,488円	291円	2.2%	CPIに一部採用済（化粧品（化粧クリーム等）、振込手数料など）

出典：総務省「平成26年全国消費実態調査（二人以上世帯）結果」より作成（購入先別に関する調査は平成26年11月に実施）

※「ネット購入」は、インターネット上で注文を行い、品物を購入またはサービスの提供を受ける形態をいう。

# 【参考5】諸外国におけるウェブスクレイピングの活用状況

	ウェブスクレイピング技術によるネット販売価格の利用		法的根拠・事前承諾の有無
	取込済の品目	研究中の品目	
オランダ	・衣料品	・航空運賃	<ul style="list-style-type: none"> <li>・ウェブサイト上のデータは誰でも利用できるもので承諾を取る必要はないが、<b>データ収集をする理由について、電子メールか書簡で企業側に知らせている。</b></li> <li>・ウェブスクレイピングについての法的根拠はないが、データ収集全体については、統計法に基づいている。</li> </ul>
イタリア	・鉄道乗車券	・航空運賃	<ul style="list-style-type: none"> <li>・ウェブスクレイピングは、鉄道運賃についてのみ、主要国営企業のサイトに行っているが、イタリア統計局長の署名入り書簡により正式に連絡し、<b>I Pアドレスを通知後行っている（ホワイトリストに入れてもらうため）。</b>また、<b>サーバーに負荷をかけないように一日のうちある時間帯にはスクレイピングしない。</b></li> <li>・法的根拠は、国家統計プログラムである（調査回答者は、必要な情報を提供する義務がある）。</li> </ul>
イギリス	—	<ul style="list-style-type: none"> <li>・食品</li> <li>・Blu-Rayディスク</li> <li>・DVD、CD</li> <li>・PC周辺機器（プリンタとルータ）</li> <li>・バック旅行 <ul style="list-style-type: none"> <li>・ラップトップ（品質調整用）</li> </ul> </li> <li>・衣料品</li> </ul>	<p>2017年10月1日付けで、ウェブスクレイピングポリシーを発効及びイギリス統計局HP内に掲載。ポリシーの内容は以下のとおり。</p> <ul style="list-style-type: none"> <li>・1年ごとに見直される。</li> <li>・CPI作成に限らず、センサスや社会調査、企業分類システムにも適用。</li> <li>・APIの利用や統計局以外での利用は適用範囲外。</li> <li>・統計局のスクレイピングに透明性等をもたせることが目的。</li> <li>・ウェブサイト所有者の最小限の負担、サイト所有者のスクレイピング中止依頼の受領、個人情報保護、法令遵守等が原則。</li> <li>・他のデータソースと比較して最もウェブスクレイピングデータが優れていると判断したときのみ利用。</li> </ul>
アメリカ	—	<ul style="list-style-type: none"> <li>・航空運賃</li> <li>・自動車燃料</li> </ul>	<ul style="list-style-type: none"> <li>・アメリカ労働統計局は<b>ウェブサイトに対するスクレイピングやAPIの利用は企業からの承諾がなければ実施できない</b>ので、フィールドスタッフが電話又は企業を訪問して口頭で承諾を取る予定。</li> <li>・フィールドマネージャーはスタッフが企業にスクレイピングやAPIの使用を説明できるように詳細な指導を行う。</li> <li>・ウェブスクレイピングに特化したポリシーは公表しておらず、秘密保護誓約書及び法を遵守している。</li> </ul>
オーストラリア	非公開	・服装、履物	<ul style="list-style-type: none"> <li>・オーストラリア統計局ではウェブサイトの利用条件及びサイトのロボットテキストファイルをプログラマーが確認し、スクレイピングすべきサイトか確認する。その後、<b>スクレイピングするサイトの小売業者に対して、まもなくスクレイピングを開始する旨を電話またはメール等で知らせている。</b></li> <li>・「プライバシー法（1998）」及び「センサス・統計法（1905）」が、ウェブスクレイピングの法的根拠であり、ウェブスクレイピングに特化したポリシーはない（サイト所有者に強制的にスクレイピングを許可させることはできない）。</li> </ul>

出典：各国HP（2019.2.13時点）、CPIに関する欧州専門家会合（2016,2018）、オタワ会議（2015,2017）、各国への問合せ（2019.7月時点）等による。