

Explaining Zipf's law by Rapid Growth

Yoshiyuki Arata

Research Institute of Economy, Trade and Industry
National Tax College

H. Yoshikawa

University of Tokyo
Policy Research Institute, Ministry of Finance

S. Okamoto

National Tax College

Ch1 Research Motivation

Zipf's law

企業売上や個人所得の分布の右テールがべき乗則に従うという統計的な性質:

$$\log \mathbb{P}(S > x) = -ax + b$$

- S : 企業売上や個人所得の対数値
- 国や年代を問わず、広く成り立つ
- 超巨大企業や超高所得者が存在する事を示唆

Question: 何故、Zipf's law は成り立つのか？その普遍的なメカニズムとは何か？

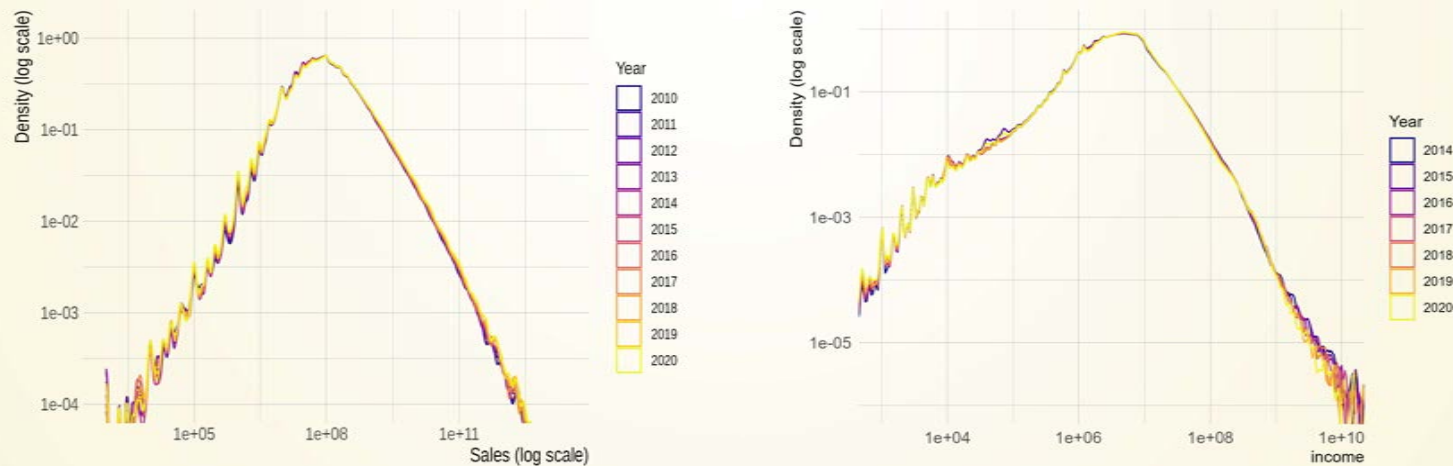


Figure: 日本の企業売上の分布 (left) と個人所得の分布 (right)

先行研究

これまでも数多くの理論モデルが提示されている

- 所得に関して: Gabaix et al. (2016 ECTA)
- 企業売上に関して: Luttmer (2011 RES)

既存モデルとデータと非整合な点 (e.g., Beare, Toda (2022 ECTA)):

- 既存モデルでは、Zipf's law は年齢別サイズ分布を集計した結果として生じる
年齢別のサイズ分布は Zipf's law に従わない
特に年齢の高い企業・個人がサイズ分布のテール部分を占める
- サイズ分布が定常分布へと収束するまでに長い時間を要する (i.e., 多くは $n \rightarrow \infty$ と仮定)
巨大企業や超高所得者になるまでに長い時間を要する

企業売上の年齢別のサイズ分布

Zipf's law はむしろ若い企業で成り立つ (e.g., 企業年齢50歳以下)

- 年齢の高い企業では Zipf's law というより、釣り鐘型の形状に近くなる
- 5歳毎にサンプルを分けても同様
年齢グループ別のサイズ分布で既に Zipf's law が成り立ち、それらのテールは共通の傾きを持つ
Zipf's law を成り立たせているのは、むしろ若い企業グループ

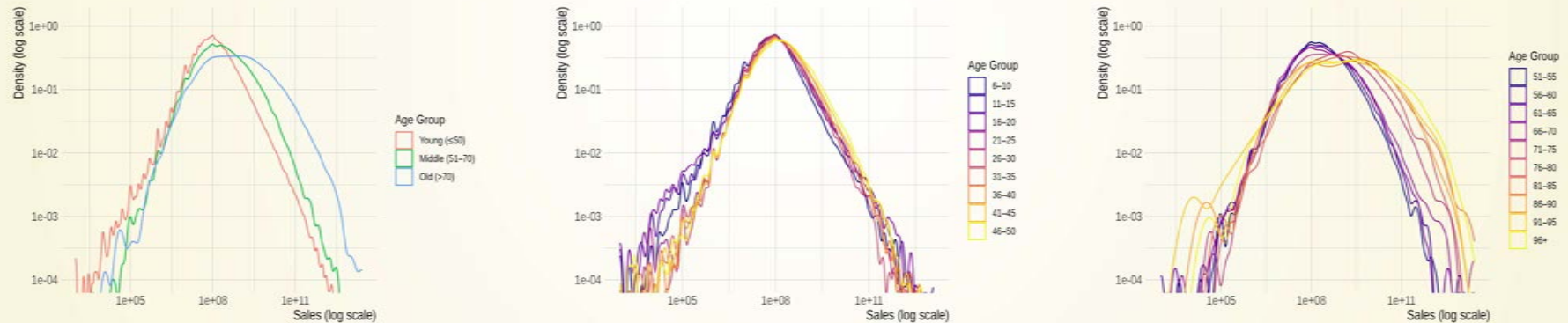


Figure: 企業売上の年齢グループ別分布。サンプルを Young (50歳以下), Middle (50-70歳), Old (70歳以上)に分解 (左図)
50歳以下のサンプルを5歳ごとに分割 (中央図)
50歳以上のサンプルを5歳ごとに分割 (右図)

個人所得の年齢別のサイズ分布

個人所得の分布についても同様なことが言える

- 5歳毎の年齢別のサイズ分布において、Zipf's law が成り立つ
若い年齢グループで既に Zipf's law が成り立つ
- 年齢が上がるにつれて平行移動しているように見える

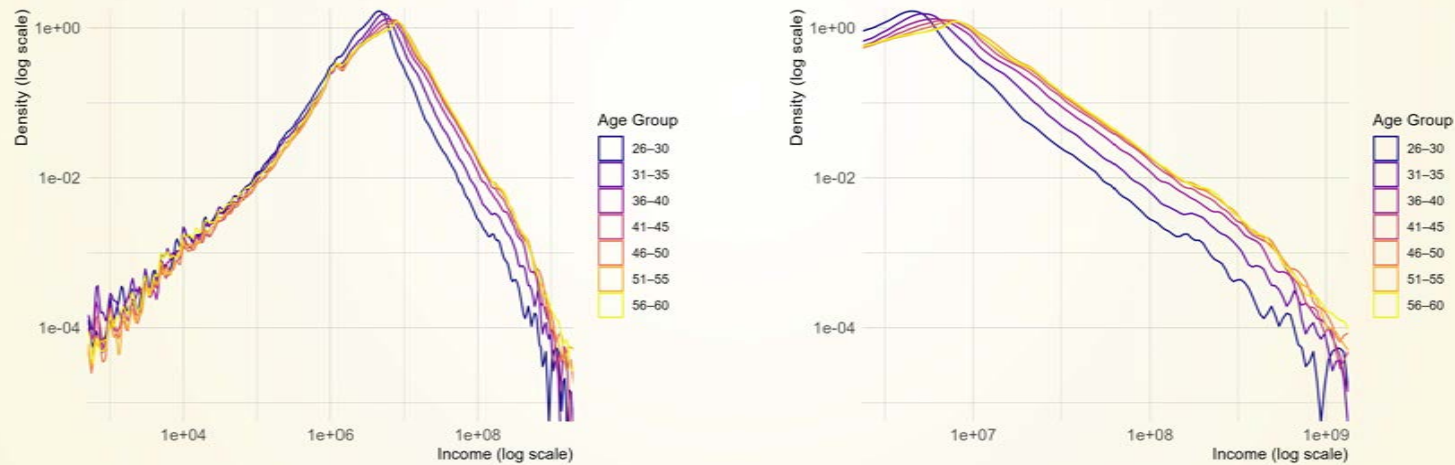


Figure: 個人所得の年齢別グループ別分布。サンプルを5歳ごとに分割 (左図)
同じ図をテール部分だけ描写 (右図)

テール確率における年齢別の寄与

全体のサイズ分布のテール $\mathbb{P}(S > x)$ における、各年齢グループが占める割合

- x の変化に対し一定、もしくは年齢の若いグループの比率が上がる
 x が大きくとも、より年齢の高い企業・個人に占められるわけではない (既存モデルの予測と異なる)

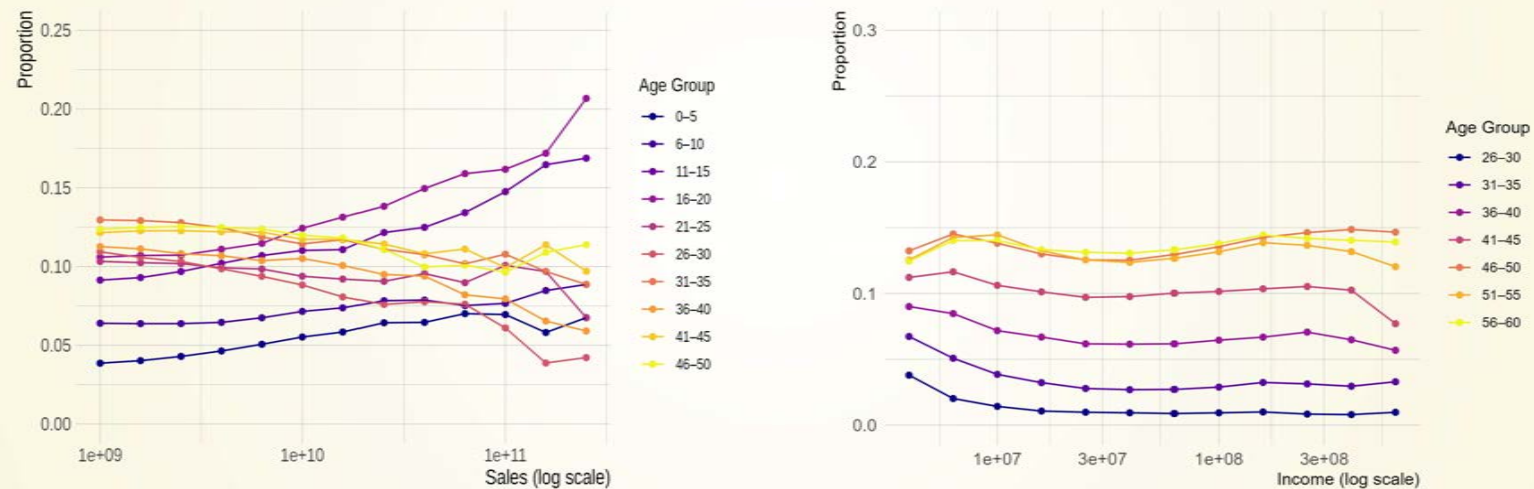


Figure: 全体のサイズ分布のテール $\mathbb{P}(S > x)$ における、各年齢グループが占める割合
企業売上のケース (左図) と個人所得のケース (右図)

本論文の目的

今回の論文の貢献:

- 既存モデルとデータとの矛盾を解決する Zipf's law の説明が必要
若いグループの年齢別サイズ分布において、どうして Zipf's law が成り立つのか？
- 企業売上と個人所得の両方に適用できる Zipf's law のメカニズムとは何か？

確率の論理で Zipf's law を説明する

Ch2 Alternative Explanation

メインアイデア

S_n を n 期の企業売上・個人所得の対数とする

$$S_n = S_0 + X_1 + X_2 + \dots + X_n$$

- 成長率 $X_1, \dots, X_n$ を独立同分布の確率変数とする
- 成長率分布が subexponential とする (i.e., 成長率分布のテールが指数関数よりも重い)
モーメント母関数 $\mathbb{E}e^{\lambda X_k}, \lambda > 0$ が存在しない

成長率分布が subexponential であれば、ジャンプによる急成長が最も典型的なプロセス

$$\mathbb{P}\left(\sum_{k=1}^n X_k > x\right) \sim n\mathbb{P}(X_k > x) \quad \text{as } x \rightarrow \infty$$

- 右辺は $X_1, \dots, X_n$ の最大値が x を超える確率 (the principle of a single big jump)

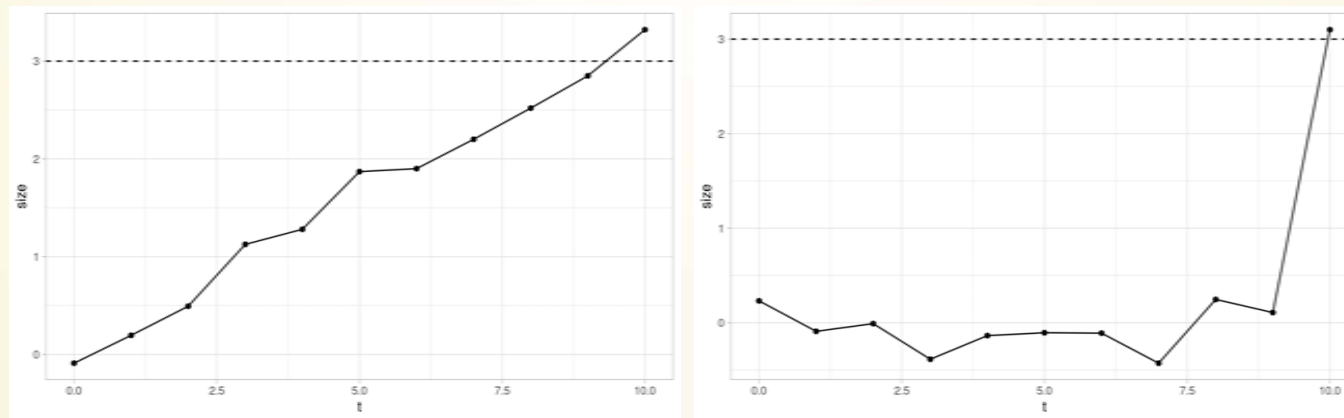


Figure: サンプルパスのイメージ図。コツコツ型の成長 (左図) とジャンプによる成長 (右図)

企業売上と個人所得のデータ

- 東京商工リサーチの日本企業のデータ (100万社)
成長率の分析では売上 4000 万円以下の企業を除く
- 税務大学校から提供された確定申告のデータ (2300万人)
個人所得の定義: 収入金額を用いる (営業収入、農業収入、不動産収入、給与収入、業務雑収入の合計)
成長率の分析では売上 400 万円以下の企業を除く

年齢の分布は両サンプルで大きく異なるが、両ケースで Zipf's law は成り立っている

- 年齢構成は Zipf's law に関係していない (年齢別で既に成り立っているため)

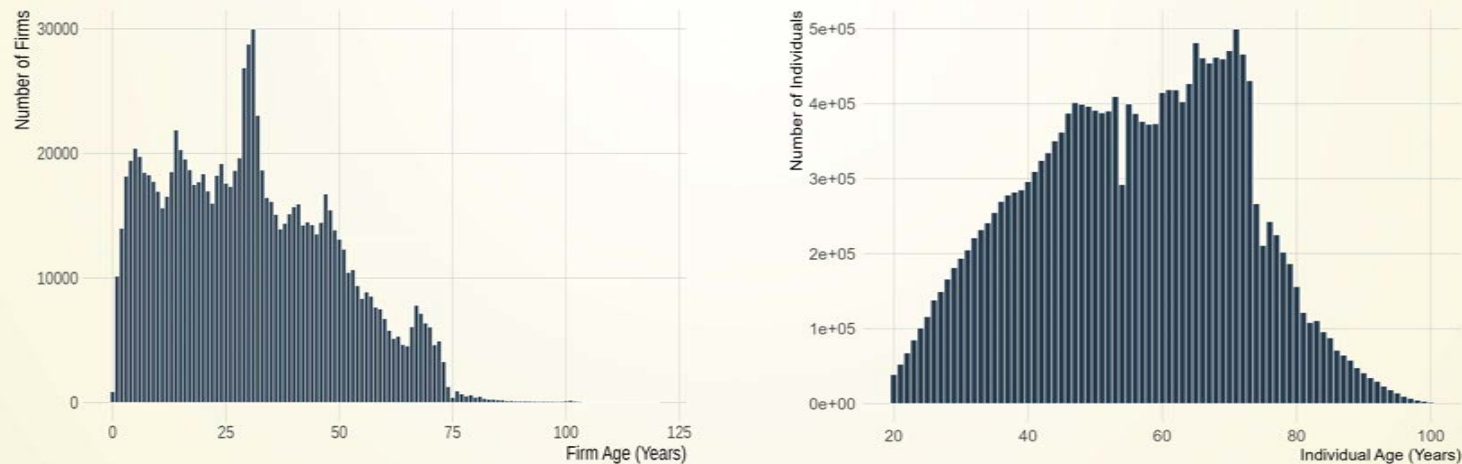


Figure: 企業・個人の年齢のヒストグラム。企業売上のケース (左図) と個人所得のケース (右図)

企業売上の成長率の分布

成長率の分布が Subexponential であることを確かめる

- 正規分布に比べて、中心部分で尖り、テール部分が重い
- 縦軸も対数で見た場合、上側にカーブしている
テール部分は指数関数よりも重い
- 複数年間の成長率分布のテールは、1年間のテールと近い傾きを持つ（Subexponential の性質）

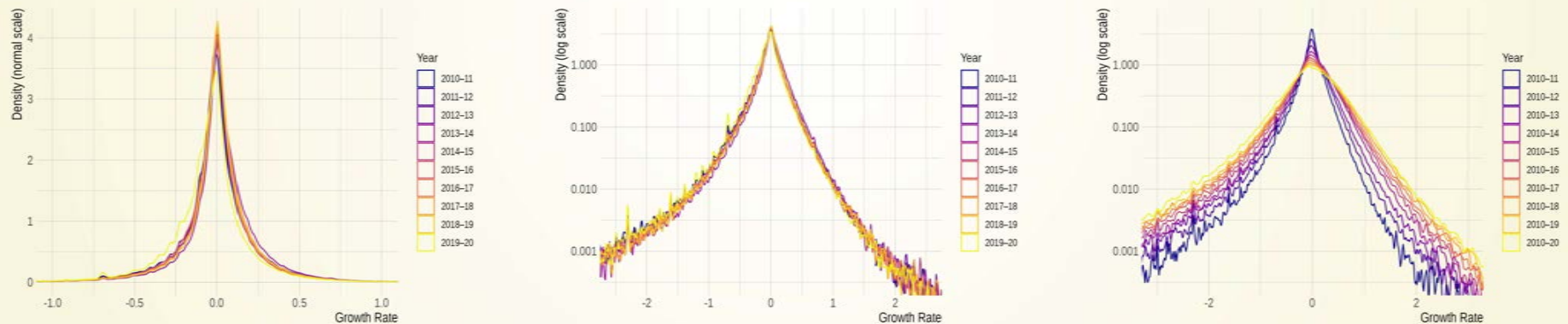


Figure: 企業売上の成長率の分布。1年間の成長率分布のノーマルスケール（左図）と log スケール（中央図）
複数年間の成長率分布の log スケール（右図）

個人所得の成長率の分布

個人所得の成長率の場合でも、同様の形状が見られる

cf. Guvenen et al. (2021 ECTA)

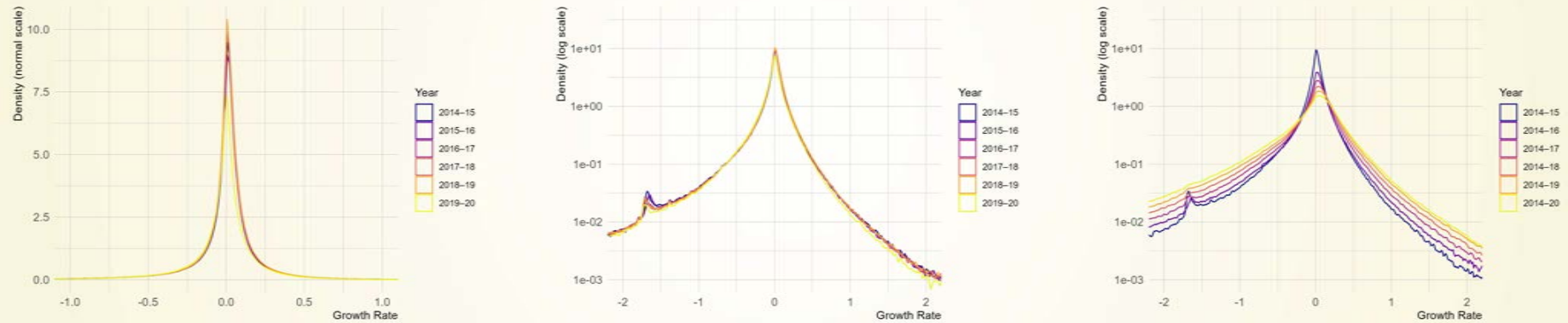


Figure: 個人所得の成長率の分布。1年間の成長率分布のノーマルスケール（左図）と log スケール（中央図）
複数年間の成長率分布の log スケール（右図）

Mean Excess Function

Mean Excess Function の定義:

$$e(u) := \mathbb{E}[X - u \mid X > u], \quad u > 0$$

- テールが指数関数ならば $e(u)$ は定数
- テールが指数関数より重いならば、 u の増加関数

企業売上・個人所得ともに u の増加関数であり、Subexponential であることを示している

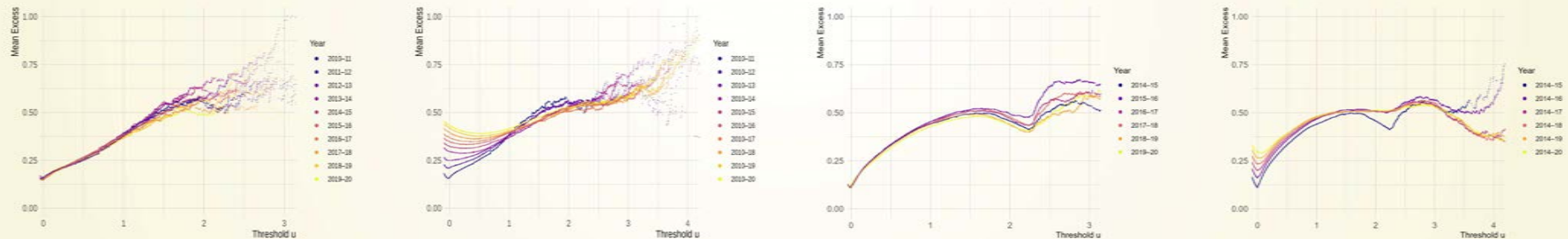


Figure: Mean excess function. 1年間の売上成長率（左図）と複数年間の売上成長率（中央左図）
1年間の所得成長率（中央右図）と複数年間の所得成長率（右図）

Ch3 Empirical Results

Tail exponent

S_n の分布の傾き a_n は、 S_0 の分布の傾きと $\sum_{k=1}^n X_k$ の分布の傾きの加重平均になる

$$S_n = S_0 + X_1 + X_2 + \dots + X_n, \quad \log \mathbb{P}(S_n > x) = -a_n x + b_n$$

- 個人所得の場合、 S_0 の分布の傾きと $\sum_{k=1}^n X_k$ の分布の傾きは 1.8 程度
その結果 S_n の傾きも 1.8
- 企業売上の場合、 S_0 の分布の傾きは 0.8、 $\sum_{k=1}^n X_k$ の分布の傾きは 1.9 程度
 n が増えるにつれて、 S_n の傾き a_n も増加する

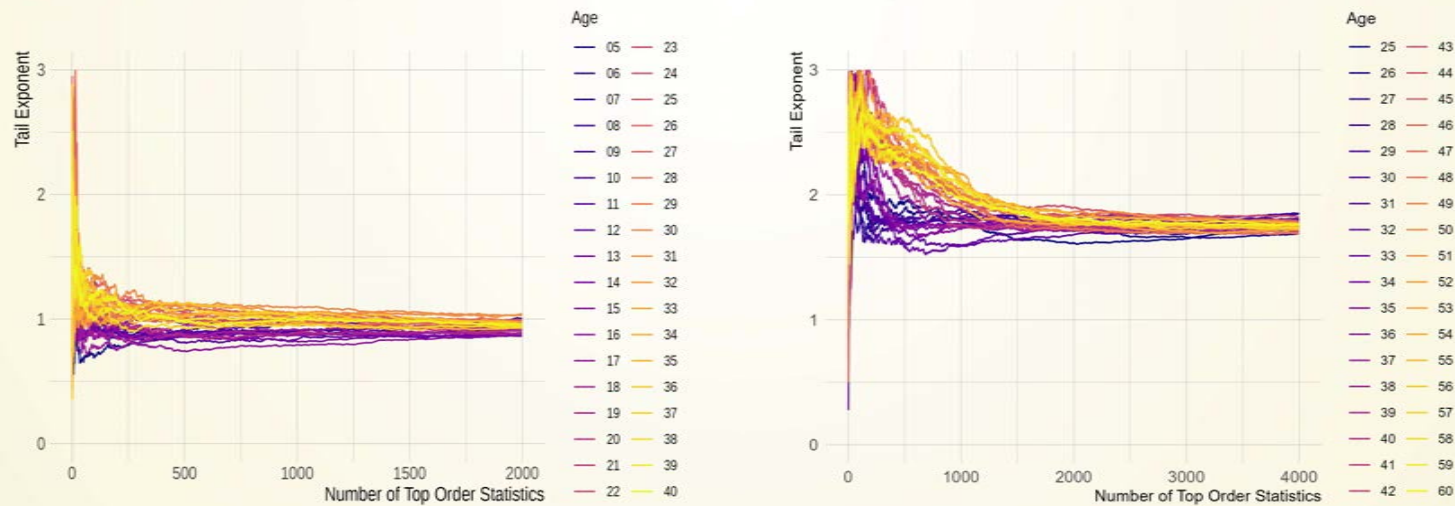


Figure: S_n の分布のテールの傾きの推定値。企業売上のケース (左図) と個人所得のケース (右図)

サンプル期間全体に対する前半期間の寄与 r

企業・個人ごとに、サンプル前半期間の成長率の全期間の成長率への寄与 r を計算

$$r = \frac{\sum_{k=1}^{n/2} X_k}{\sum_{k=1}^n X_k}$$

- $\sum_{k=1}^n X_k > u$ の高成長のサンプルに絞ってヒストグラムを求める
- u が小さいと、 $r = 1/2$ が最も生じやすい
- u が大きくなると、 $r = 0, 1$ でスパイクを持つ
全期間の高い成長率は、前半期間か後半期間のどちらか一方の高成長で達成している

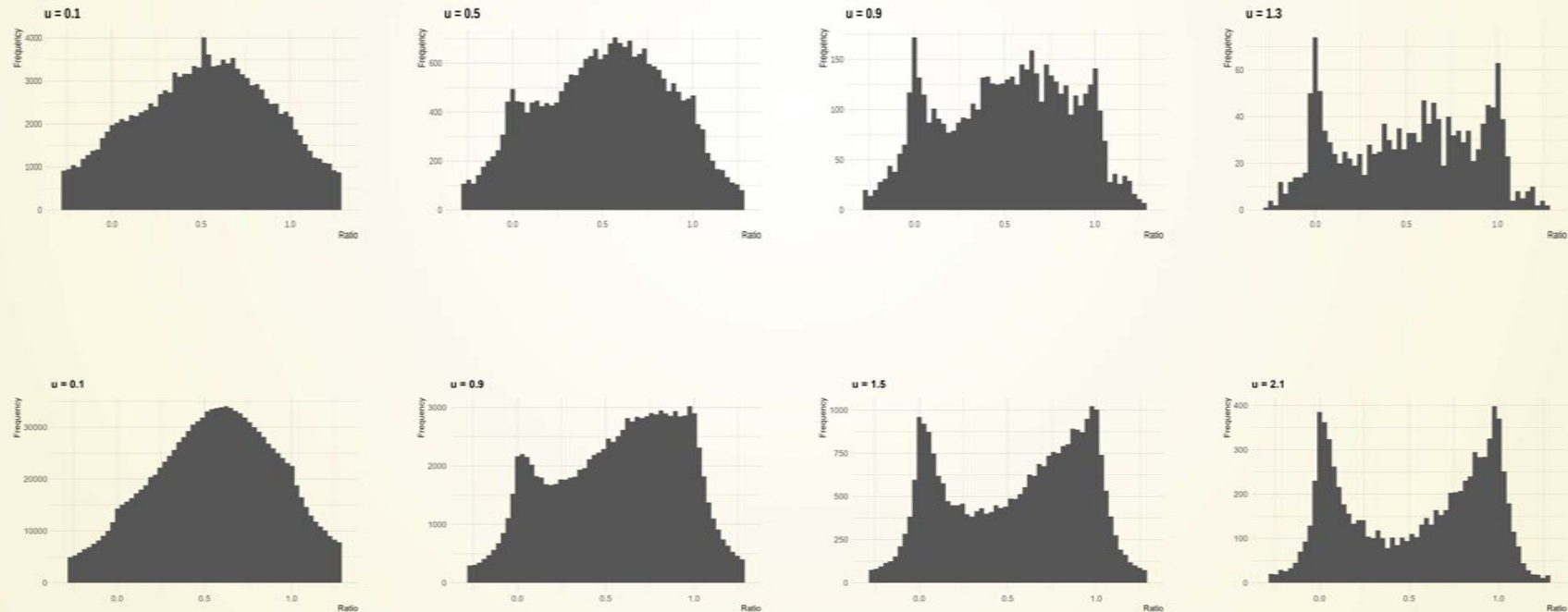


Figure: $\sum_{k=1}^n X_k > u$ で条件付けた r のヒストグラム。企業売上のケース, $n = 2$ (上部) と個人所得のケース, $n = 6$ (下部)

個人・企業ごとのサンプル期間における成長率の最小値

n 期間で高い成長率 x を達成したグループ (A) とそうではないグループ (B) において

- ジャンプによる高成長であれば、ジャンプ以外の成長率は、A と B で差が無いはず
- 平均的な成長によるものであれば、 n 個の成長率全てが x/n だけ正方向にドリフトするはず

ジャンプ以外の期の成長率の分布が、グループ A と B で異なるかを調べる

- 企業・個人ごとにサンプル期間の成長率の最小値を求め、グループごとにその分布を比較
- グループ間で成長率の最小値の差が小さい（ジャンプ以外の期は成長率分布に差が無い）
ジャンプによる成長を支持する結果

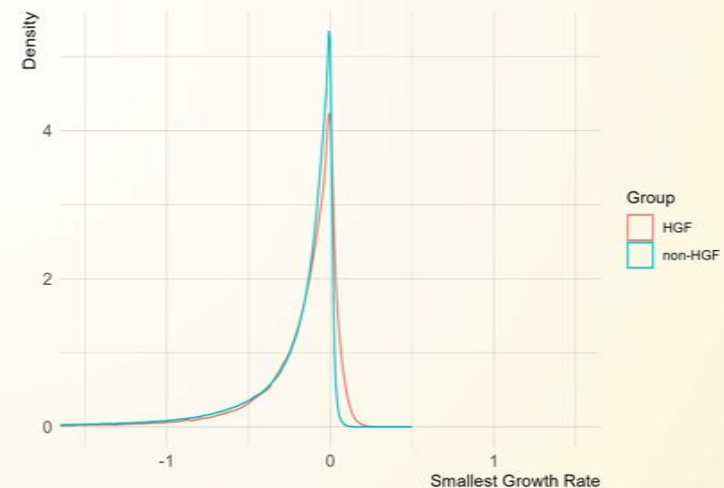
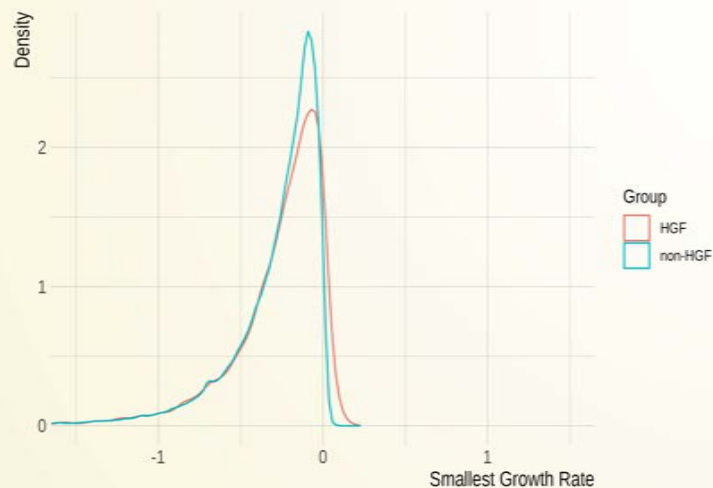


Figure: 企業・個人ごとの期間中の最小成長率の密度関数推定。企業売上のケース (左図) と個人所得のケース (右図)

Ch4 Conclusion

Zipf's law の既存モデルに代わる新たな説明

- Zipf's law は年齢別のサイズ分布で既に成り立つ
Zipf's law は成長率分布と初期サイズの分布のテールの重さで決まる
- Zipf's law はジャンプによる成長プロセスを反映
短期間でテールに到達し得る ($n \rightarrow \infty$ のような定常性の仮定は必要としない)

Appendix

複数タイプの agent が存在するモデル (e.g., Beare, Toda (2022 ECTA)) との違い

- (1) Zipf's law は異なる年齢の agent を集計した結果ではなく、年齢別のサイズ分布で成立する
特に、若い年齢グループで Zipf's law が成り立つ
企業売上では、年齢の高いグループのほどガウス分布に近づくことも確認
- (2) 年齢別サイズ分布のテール指数は、成長率分布と初期サイズ分布のテール指数によって決まる
企業売上・個人所得ともに Pareto tail を持つが、テール指数は異なる
この違いは成長率分布と初期サイズ分布のテール指数の違いに対応
- (3) n 期間の高成長が、ジャンプによって決まる
我々の理論: n 期間の高成長は1つのジャンプによって達成
既存モデル: n 個すべてが平均的に高めの成長率の累積によって達成
データでは、ジャンプ期以外の成長率は正方向にドリフトしていない

理論的差異: 成長率のモーメント母関数 $\mathbb{E}e^{\lambda X_k}, \lambda > 0$ が存在するか否か、これにより Large Deviation は根本的に異なる