# 演習: 不平等の測定

計量経済学各論 (開発経済の計量分析)

2016 年 7 月 14 日版

## 目次



## 1 目的

Penn World Table 9.0 の国内支出 (Domestic Absorption) を使って，国間の不平等について以下を計算する．

1. ローレンツ曲線
2. ジニ係数
3. タイルのエントロピー指数
4. タイルのエントロピー指数を先進国と開発途上国に関する地域間効果と地域内効果に分解する．

## 2 データ

授業の Web サイトにおける「演習と課題」の「開発指標」のセクションから Penn World table 9.0 のデータを収納した SQLite データベース `pwt90.db` を作業ディレクトリにダウンロードする．SQLite は，パブリックドメインのリレーショナルデータベース管理システム (RDBMS) である．課題 1 で利用した 2 つのパッケージ `dplyr` と `sqldf` がインストールされていれば R から SQLite を使うことができる．これら 2 つのパッケージをインストールしていないものは，ただちにインストールすること．

`pwt90.db` には以下の 2 つのテーブルが登録されている．

| テーブル名 | 内容 |
|---|---|
| pwt | Penn World Table 9.0 本体 |
| country | PWT9.0 に含まれる国情報．国連世界人口推計の国情報に基づく． |

## 3　必要なパッケージの読み込み

以下のパッケージが必要になる．インストールしていないものがあればインストールしておく．人口の演習で使った sqldf パッケージがインストールしてあればそのときに RSQLite パッケージが同時にインストールされている．tidyr パッケージは，はじめての利用かもしれない．

```
library(RSQLite)
library(dplyr)
library(tidyr)
library(ggplot2)
library(reldist)
```

## 4　データの読み込みと整理

SQLite には SQL でデータベースに問い合わせてデータを取得する．pwt テーブルからは，国コード (ISO3)，年次 (year)，人口 (pop)，国内支出 (cda) のみを取り込む．country テーブルについては全てのデータを取り込む．

```
driv <- dbDriver("SQLite")
dbname <- "pwt90.db"
con <- dbConnect(driv, dbname)
pwt <- dbGetQuery(con, "SELECT ISO3, year, pop, cda FROM pwt")
country <- dbGetQuery(con, "SELECT * FROM country")
dbDisconnect(con)
```

```
## [1] TRUE
```

1 行づつ説明してみよう．

1. R からは，SQLite のみならず，MySQL や PostgreSQL など様々な RDBMS に接続することができる．どの RDBMS にも DBI パッケージを介して同じ方法でアクセスすることができる．ここでは SQLite を使うことを driv <- dbDriver(RSQLite) で指定している．
2. 使用するデータベース名を dbname に代入する．
3. dbConnect 関数で SQLite の pwt90.db に接続する．以後，返り値の con を介してデータベースと通信する．
4. dbGetQuery 関数は，接続先 con の SQL を問い合わせてデータを取得する．2 番めの引数に SQL を与える．SELECT ISO3, year, pop, cda FROM pwt は，pwt テーブルから ISO3(国コード)，year(年次)，pop(人口)，cda(国内支出) を取得することを SELECT 文で命令している．データは，データフレームとして pwt に保存される．
5. 同様に，country テーブルから全てのデータを取得する．結果は，データフレーム country に入る．

country テーブルには各国を所得グループに分類するための変数 IncomeGrp がある．ImcomGrp によって各国は高所得国 (HIC)，高位中所得国 (HMIC)，低位中所得国 (LMIC)，低所得国 (LIC) に分類することができるが，今回の演習では各国を開発途上国と先進国の 2 地域に分類する．ただし，高所得国を先進国とし，それ以外を開発途上国とする．そのために country データフレームに開発途上国であることを示す Developing という新しいダミー変数を作り，開発途上国であれば Developing = 1 とし，先進国であれば Developing = 0 とする．

```
country <- mutate(country, Developing = ifelse(IncomeGrp == "HIC", 0, 1))
```

pwt データフレームに counyry の所得グループ変数 (IncomeGrp と Developing) を国コード (ISO3) をキーにしてマージする．まず，マージする前の pwt テーブルの内容確認すると，

```
head(pwt)
```

```
##   ISO3 year pop cda
## 1  ABW 1950  NA  NA
## 2  ABW 1951  NA  NA
## 3  ABW 1952  NA  NA
## 4  ABW 1953  NA  NA
## 5  ABW 1954  NA  NA
## 6  ABW 1955  NA  NA
```

となっており，ISO3 に対応する国がどの所得グループに属するのかわからない．country テーブルから，ISO3，ImcomeGrp，Developing だけを選んで，ISO3 をキーにして pwt とマージする．

```
pwt <- left_join(pwt, select(country, ISO3, IncomeGrp, Developing))
```

```
## Joining by: "ISO3"
```

結果を確認すると，

```
head(pwt)
```

```
##   ISO3 year pop cda IncomeGrp Developing
## 1  ABW 1950  NA  NA       HIC          0
## 2  ABW 1951  NA  NA       HIC          0
## 3  ABW 1952  NA  NA       HIC          0
## 4  ABW 1953  NA  NA       HIC          0
## 5  ABW 1954  NA  NA       HIC          0
## 6  ABW 1955  NA  NA       HIC          0
```

となって，ABW(Aruba，アルバ，カリブ海諸国) は，高所得国 (ImcomeGrp = HIC) で先進国 (Developing = 0) に属することがわかるようになった．

PWT 9.0 に含まれる国で国連世界人口推計では所得グループが設定されていない国があるので，それらは除外しておく．さらに，国内支出が利用できないレコード (cda = NA) も除外する．また，タイルのエントロピー指数では cda の対数を取るので，cda が非正値のレコードも除外しておく．

```
pwt <- filter(pwt, !is.na(Developing) & !is.na(cda) & cda > 0)
```

PWT 9.0 には 1950 年から 2014 年までの 65 年間のデータが収録されている．2014 年では 178 カ国のデータが利用できるが，時代を遡るほどデータの遡及は困難であり，1950 年では 55 カ国のデータしか利用できない．データが利用可能な国のうち開発途上国の割合は，1950 年で 49.1%，データが利用可能な国のうち開発途上国の割合は，2014 年で 64% である．不平等指標の時系列推移を見る場合には，この点に注意を払うべきである．

## 5　不平等測定のための準備

サイズ $N$ の有限母集団 (総世帯数) から復元抽出されたサンプルサイズ $n$ の標本調査における世帯 $i$ の所得 (あるいは消費) を $Y_i$ $(i = 1, \cdots, n)$ と書く．世帯 $i$ が抽出される確率が $p_i$ のとき

$$N_i = \frac{1}{np_i}, \quad i = 1, \cdots, n \tag{1}$$

をウエイトと呼ぶ． $np_i$ は，世帯 $i$ が調査に現れる期待値を示すが，母集団サイズ $N$ に比してサンプルサイズ $n$ が小さいとき，世帯 $i$ が重複して抽出される確率は極めて小さいと考えられるので，$np_i$ を世帯 $i$ の抽出確率と近似しても差し支えない．このとき $N_i$ は，世帯 $i$ で代表される世帯の母集団における世帯数を近似する[*1]． すなわち，$n_i$ を世帯固有の膨らませ率と考えることができる．

$$\hat{N} = \sum_{i=1}^{n} N_i \tag{2}$$

は母集団サイズ (総世帯数) の不偏推定量である．そして，母集団全体の総所得 (総消費) に関する不偏推定量は次のように与えられる[*2]．

---

[*1] 単純無作為抽出の場合 $\pi_i = 1/N$ で全ての世帯で共通になる．しがって $N_i = N/n$ は，標本から母集団へ膨らませる膨らませ率 (inflation factor) である．世帯 $i$ が調査に現れる現れないを示す確率変数を $T_i$ とする．$T_i$ はベルヌーイ事象で，次の確率質量関数を持つ．

$$f(t_i) = \begin{cases} p_i, & t_i = 1 \\ 1 - p_i, & t_i = 0 \\ 0, & \text{それ以外} \end{cases}$$

$T_i$ の期待値は，

$$\mathbb{E}[T_i] = 1 \times p_i + 0 \times (1 - p_i) = p_i$$

サンプルサイズ $n$ の表標本を $T_{i:1}, T_{i:2}, \cdots, T_{i:n}$ とする．$T_{i:k}$ $(k = 1, \cdots, n)$ が取る値は 1 または 0 であるから世帯 $i$ が調査に現れる回数は $\sum_{k=1}^{n} T_{i:k}$ の期待値は，

$$\mathbb{E}\left[\sum_{k=1}^{n} T_{i:k}\right] = \sum_{k=1}^{n} p_i = np_i$$

[*2]

$$\hat{N} = \sum_{i=1}^{n} N_i = \sum_{i=1}^{N} T_i N_i$$

2 番目の和は母集団全体について取っているので，$N_i$ は確率変数ではなく，$T_i$ が確率変数である．したがって，

$$\mathbb{E}[\hat{N}] = \sum_{i=1}^{N} \mathbb{E}[T_i] N_i = \sum_{i=1} np_i \frac{1}{np_i} = \sum_{i=1}^{N} 1 = N$$

となる．$\mathbb{E}[\hat{Y}] = Y = \sum_{i=1}^{N} Y_i$ も同様に計算できる．

$$\hat{Y} = \sum_{i=1}^{n} N_i Y_i \tag{3}$$

以上より，標本調査の個票または階級別データとして階級の標本平均と世帯数が与えられている場合，所得分布の観察値は，$\{n_1 y_1, n_2 y_2, \cdots, n_n y_n\}$ である．ここで，$n_i$，$y_i$ はそれぞれ確率変数 $N_i$，$Y_i$ の実現値である．また，世帯の所得 (消費)$y$ は，$y_1 < y_2 < \cdots < y_n$ の順序で並んでいる (匿名性の公理)．

世界銀行の LSMS(Living Standard Measurement Study) などで提供されている標本調査の個票では，調査世帯の構成人員数が報告されていることが多い．世帯 $i$ の人員数を $m_i$ とすれば，$m_i n_i$ は世帯 $i$ タイプの家計に属する人口の推定値を与える．そして世帯内で所得 (消費) が均等に配分されていると仮定すれば，一人当たり所得 (消費) は $y_i/m_i$ で与えられる．よって改めて，$y_i$ を $y_i/m_i$，$n_i$ を $m_i n_i$ とすれば，世帯ベースではなく個人ベースの不平等を測定できることになる．

一方で，PWT においては各年次のマクロの国内支出額 $x$ が国際間クロスセクションデータ $\{x_1, x_2, \cdots, x_n\}$ として与えられている．このとき，人口の一人一人が同じ支出水準であることを仮定して，また $n_i$ を国 $i$ の人口として，$x_1/n_1 < x_2/n_2 < \cdots < x_n/n_n$ の順序で並んでいる．

## 6　ローレンツ曲線

次の関数は，ローレンツ曲線を描く関数 LorenzCurve の定義である．

```
LorenzCurve <- function(lcdata, ylab = NULL, glab = NULL, label = NULL) {
    if(is.null(lcdata$group)) bygroup <- FALSE
    else bygroup <- TRUE
    if(bygroup) {
        lcdata <- lcdata %>%
            group_by(group) %>%
                filter(!is.na(y)) %>%
                arrange(y) %>%
                mutate(L = cumsum(n*y) / sum(n*y), p = cumsum(n) / sum(n))
    } else {
        lcdata <- lcdata %>%
            filter(!is.na(y)) %>%
            arrange(y) %>%
            mutate(L = cumsum(n*y) / sum(n*y), p = cumsum(n) / sum(n))
    }
    if(bygroup) lc <- ggplot(lcdata, aes(x = p, y = L, colour = factor(group)))
    else lc <- ggplot(lcdata, aes(x = p, y = L))
    lc <- lc + geom_line() +
        annotate("segment", x = 0, xend = 1, y = 0, yend = 1) +
        labs(x = "Cumurative Distribution of Population",
             y = ylab, colour = glab)
    if(bygroup)
        lc <- lc + theme(legend.position=c(0,1),
```

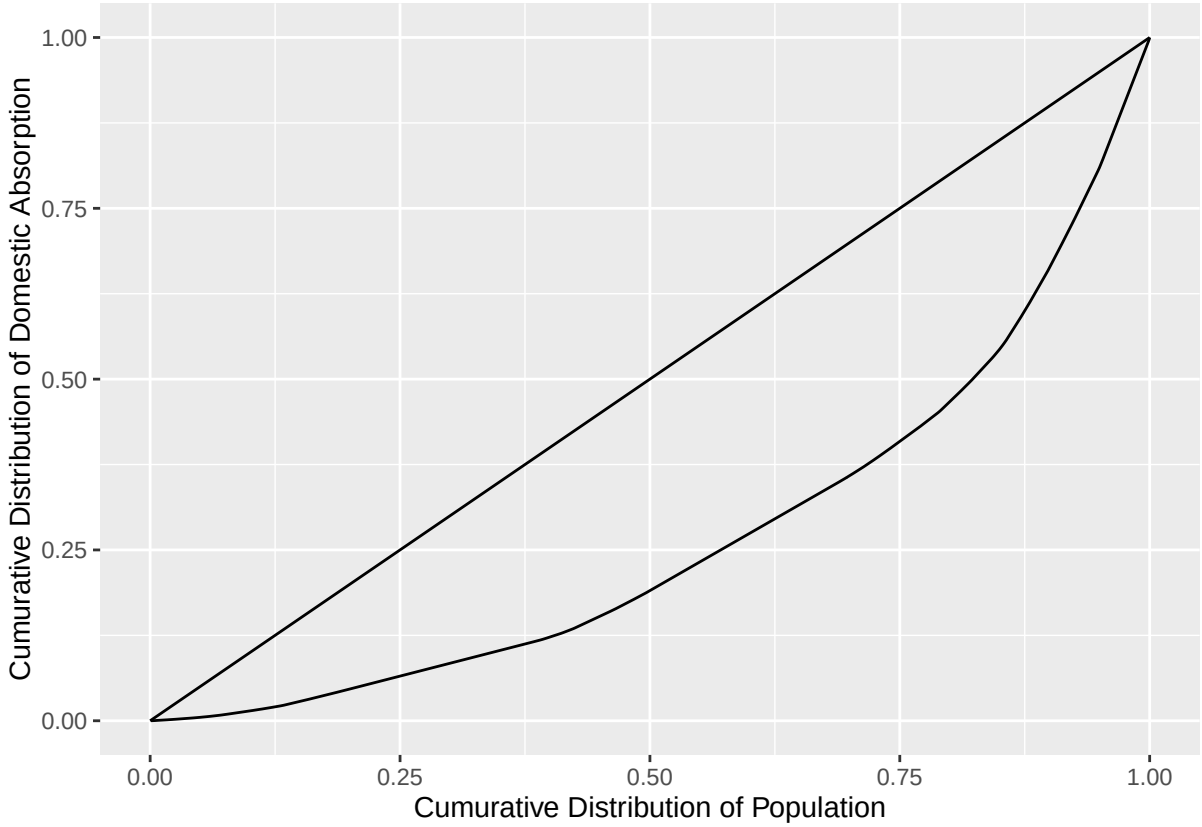
```
                          legend.justification=c(0,1))
    if(!is.null(label))
        lc <- lc + scale_colour_discrete(labels=label)
    lc
}
```

LorenzCurve には次の 3 つの変数を与える．

| 変数 | 内容 |
|---|---|
| lcdata | y，n，group の 3 変数からなるデータフレーム．group はオプショナルである．y と n は，5 節のそれぞれ $(y_1, y_2, \cdots, y_n)$，$(n_1, n_2, \cdots .n_n)$ に対応している．LorenzCurve では，y の小さい順にデータを並べ替え，横軸に n の累積分布，縦軸に n*y の累積分布をとってローレンツ曲線を描いている．group が与えられるとグループごとにローレンツ曲線を描く．たとえば，地域，年次などである．group は文字列でも数値でもよい． |
| ylab | $y$ 軸のラベル．文字列．オプショナル． |
| glab | グループ名のラベル．文字列．オプショナル． |
| label | 凡例のラベル．文字列ベクトル．オプショナル．label が与えられないときは，‘group‘の値が使われる． |

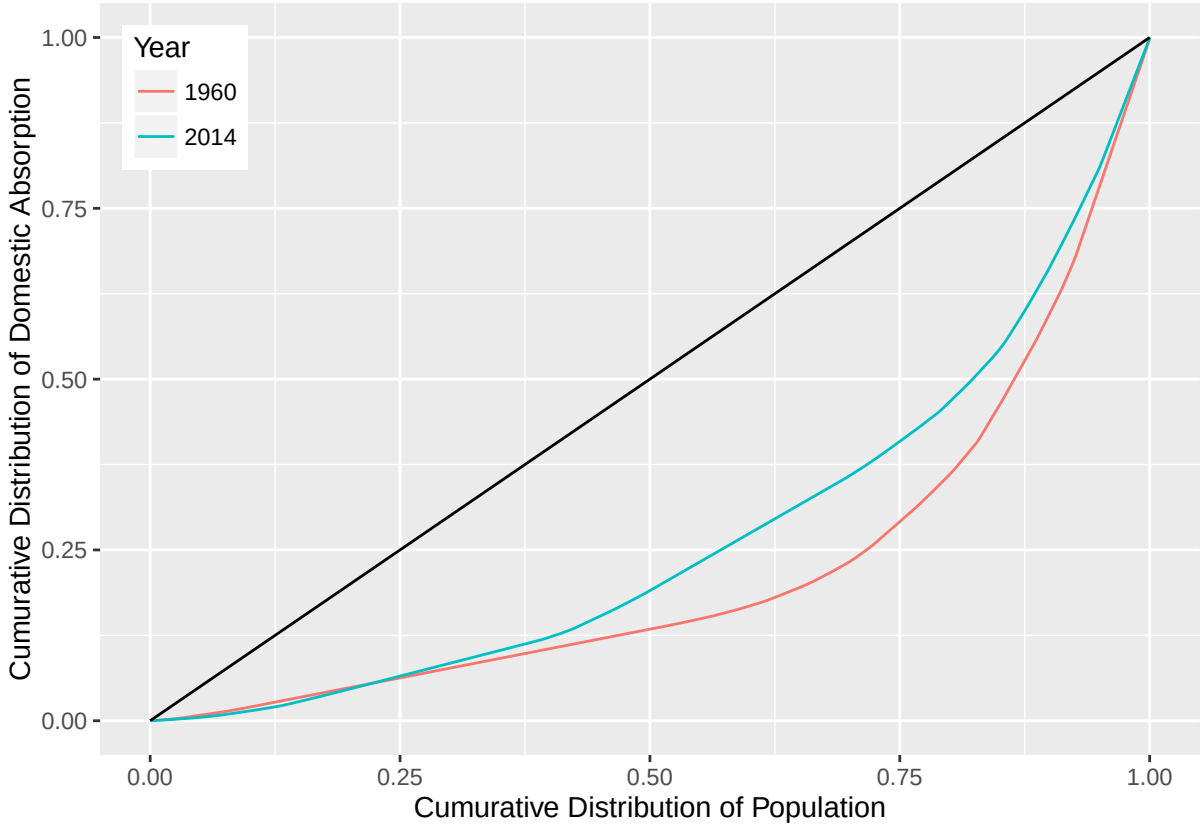
最初に 2014 年のローレンツ曲線を単独で描いてみよう．pwt から 2014 年のデータを filter し，y に一人当たりの国内支出 cda / pop，n に人口 pop を与えている．y と n だけのデーターフレームに select し，それをパイプラインで LorenzCurve に渡している．

```
filter(pwt, year == 2014) %>%
mutate(y = cda / pop, n = pop) %>%
select(y, n) %>%
LorenzCurve(ylab = "Cumurative Distribution of Domestic Absorption") -> lc2014
print(lc2014)
```




次に，1960 年と 2014 年のローレンツ曲線を重ね書きしてみよう．pwt から 1960 年と 2014 年のデータを filter していることと，group に年次 year を与えていることが，上と違っている．

```
filter(pwt, year %in% c(1960, 2014)) %>%
mutate(group = year, y = cda / pop, n = pop) %>%
select(group, y, n) %>%
LorenzCurve(ylab = "Cumurative Distribution of Domestic Absorption",
            glab = "Year") -> lc
print(lc)
```




## 7　ジニ係数

上の図では，1960 年のローレンツ曲線が 2014 年の曲線よりもほとんどの点で外側にあり，明らかに 45 度線とローレンツ曲線の間の面積が大きいから，1960 年の方が 2014 年より不平等度が大きい．ここでは 45 度線とローレンツ曲線の間の面積をジニ係数によって計算して，その時系列推移を確かめてみる．ジニ係数は，reldist パッケージの gini 関数で計算することができる．gini 関数の第 1 引数は LorenzCurve の y，第 2 引数は n に対応する．

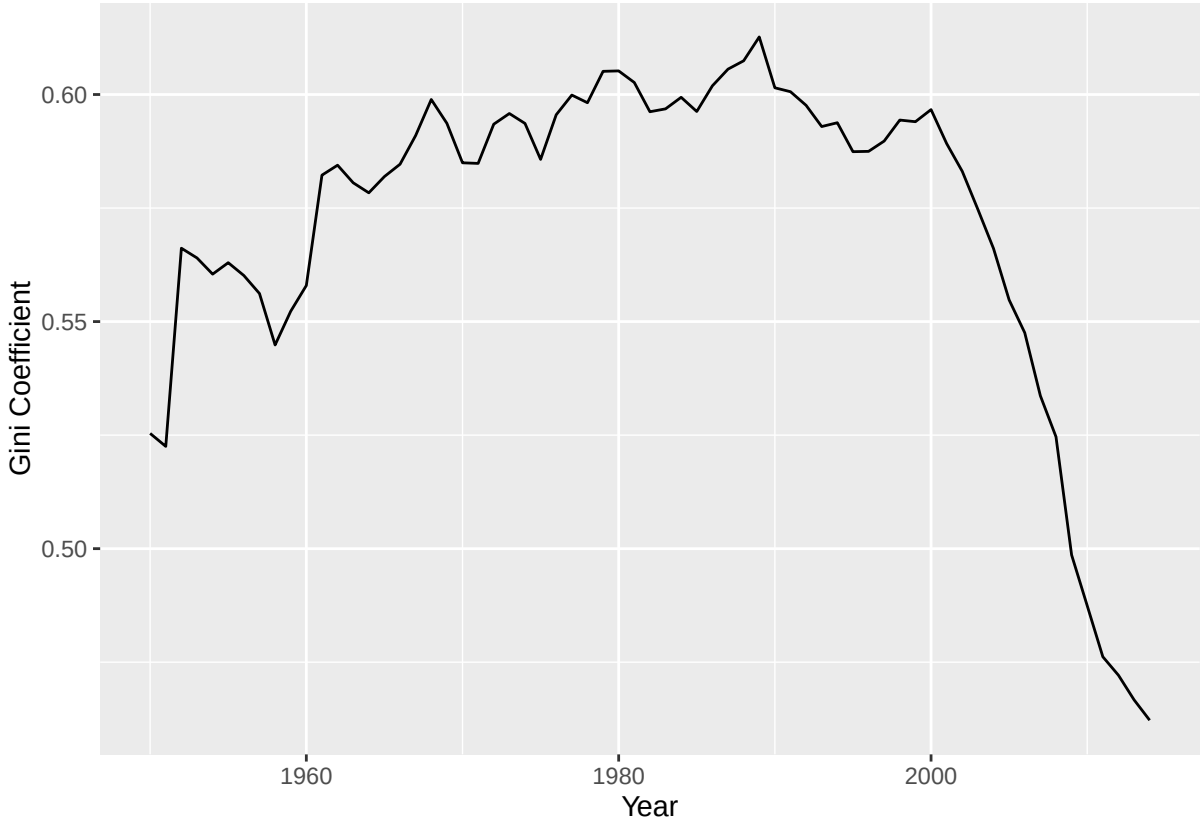
以下では，group_by(year) によって時系列でじに係数を計算している．

```
group_by(pwt, year) %>%
    summarize(G = gini(cda/pop, pop)) -> Gini
head(Gini)
```

```
## Source: local data frame [6 x 2]
##
##    year         G
##   (dbl)     (dbl)
## 1  1950 0.5253772
## 2  1951 0.5225216
## 3  1952 0.5661558
## 4  1953 0.5640101
## 5  1954 0.5604465
## 6  1955 0.5629676
```

ジニ係数の時系列推移をグラフで図示してみる．1990 年までは世界の不平等は拡大してきたが，2000 年以降急激に縮小していることがわかる．

```
ggplot(Gini, aes(x = year, y = G)) +
   geom_line() +
   labs(x = "Year", y = "Gini Coefficient") -> plotGini
print(plotGini)
```




## 8　タイルのエントロピー指数

タイルのエントロピー指数は次のように定義される．

$$T = \sum_{i=1}^{n} w_{y:i} \log\left(\frac{w_{y:i}}{w_{n:i}}\right), \quad w_{y:i} = \frac{n_i y_i}{\sum_{k=1}^{n} n_k y_k}, \quad w_{n:i} = \frac{n_i}{\sum_{k=1}^{n} n_k} \tag{4}$$

次の TheilEntropy がタイルのエントロピー指数を計算する関数である．y と n の定義は，これまでと同じである．

```
TheilEntropy <- function(y, n) {
    i <- !is.na(y) & y > 0
    y <- y[i]
    n <- n[i]
    wy <- n*y / sum(n*y)
```

```
    wn <- n / sum(n)
    sum(wy * log(wy / wn))
}
```

ジニ係数と同様にタイルのエントロピー指数を時系列で計算してみる．

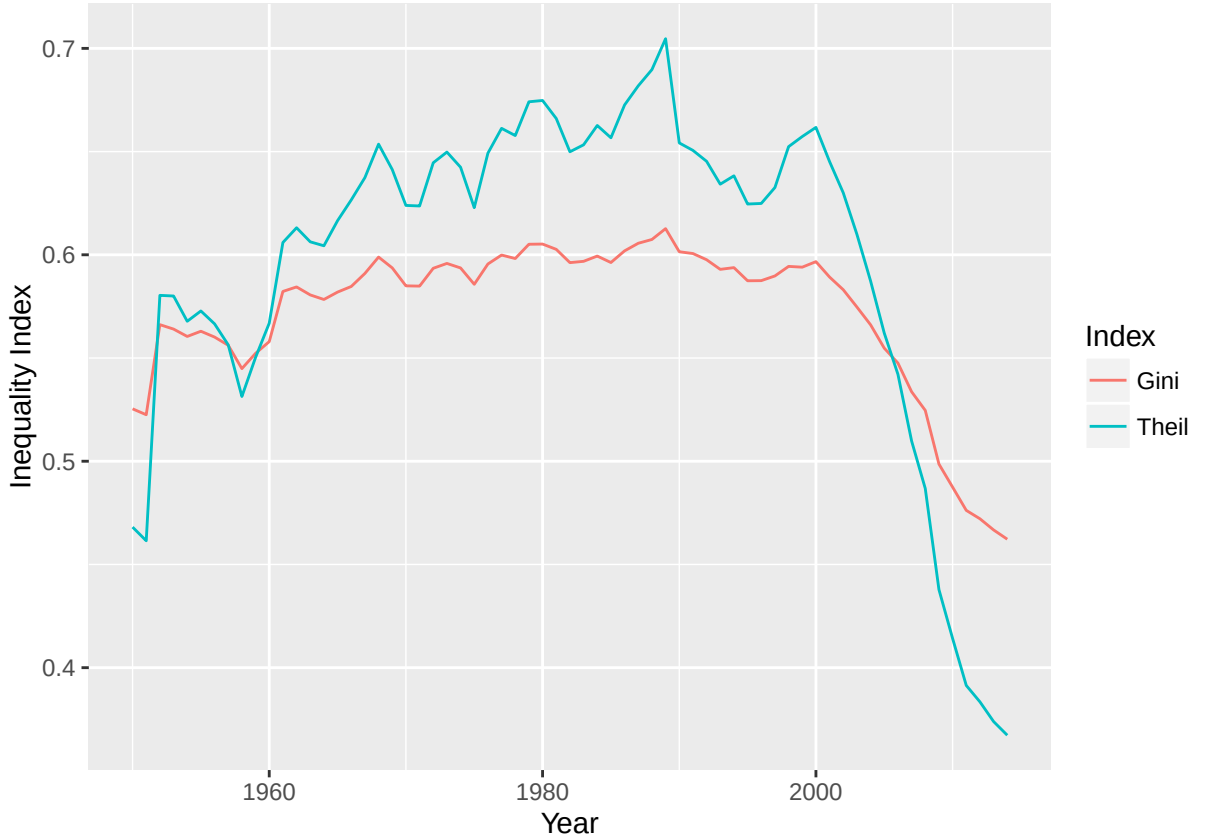
```
group_by(pwt, year) %>%
    summarize(T = TheilEntropy(cda / pop, pop)) -> Theil
```

ジニ係数とタイルのエントロピー指数の時系列推移を重ねて図示してみよう．それぞれの計算結果 Gini と Theil を年次 (year) をキーにして結合し，tidyr パッケージの gather 関数をで ggplot のデータ形式であるロング形式に変換している．

```
left_join(Gini, Theil) %>%
rename(Gini = G, Theil = T) %>%
gather(Index, val, -year) %>%
ggplot(aes(x = year, y = val, colour = Index)) +
    geom_line() +
    labs(x = "Year", y = "Inequality Index") -> plotGiniTheil
```

```
## Joining by: "year"
```

```
print(plotGiniTheil)
```




2 つの指標のトレンドは同じであるが，タイルのエントロピー指数の方が変化が大きい．これはタイルのエントロピー指数が，所得の低い方の変化に敏感に反応するという性質による．

## 9　タイルのエントロピー指数の分解

世帯番号からなる集合 $\mathcal{N} = \{1, 2, \cdots, n\}$ を互いに排反で空でない $l$ 個の集合 $\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_m$ に分解する．グループ $s$ のタイルのエントロピー指数を次のように計算する．

$$T_s = \sum_{i \in \mathcal{N}_s} w_{y:i} \log \left( \frac{w_{y:i}}{w_{n:i}} \right), \quad s = 1, \cdots, l \tag{5}$$

全サンプルにしめるグループ $s$ の所得 (消費) と世帯数の構成比以下のように定義する．

$$w_y^s = \frac{\sum_{i \in \mathcal{N}_s} n_i y_i}{\sum_{k=1} n_k y_k}, \quad w_n^s = \frac{\sum_{i \in \mathcal{N}_s} n_i}{\sum_{k=1} n_k}, \tag{6}$$

各グループの不平等指数 $T_s$ を所得構成比でウェイトした指標をグループ内不平等 (within group inequality) と呼ぶ．

$$T_W = \sum_{s=1}^{l} w_y^s T_s \tag{7}$$

また，グループごとに集計された構成比 $w_y^s$ と $w - s_n$ で計算される不平等指数をグループ間不平等 (between group inequality) と呼ぶ．

$$T_B = \sum_{s=1}^{n} w_y^s \log \left( \frac{w_y^s}{w_n^s} \right) \tag{8}$$

そして，タイルのエントロピー指数はグループ間不平等とグループ内不平等の和に完全に分解することができる．

$$T = T_B + T_W \tag{9}$$

全地域を先進国と開発途上国の 2 つのグループに分割して，タイルのエントロピー指数を先進国と開発途上国の地域間不平等とそれぞれの地域の地域内不平等に分解する．

地域内不平等の計算にはウェイトにする国内支出の構成比が必要なので，最初にその計算を行う．国内支出の世界計にしめる国別の構成比を計算しておけば，それを後に地域別に集計すれば地域別の構成比を得ることができる．もちろん構成比は年ごとに計算しなければならない．

```
group_by(pwt, year) %>%
    mutate(w = cda / sum(cda)) -> pwt
```

次に地域別のエントロピー指数を計算し (2 行目)，地域間不平等を計算するために国内支出 (cda)，人口 (pop)，国内支出の構成比 (w) を地域別に集計する (3, 4, 5 行目)．年別，地域別に計算するために year，Developing の順で group_by している (2 行目)．結果を TRegion に保存する (5 行目)．

```
group_by(pwt, year, Developing) %>%
    summarize(T = TheilEntropy(cda / pop, pop),
              cda = sum(cda),
              pop = sum(pop),
              w = sum(w)) -> TRegion
```

最後に，地域間不平等と地域内不平等の分解を行う．地域別に集計されたデータで地域間不平等を計算し (TB，2 行目)，地域別不平等指数を国内支出で加重平均して地域内不平等をを計算している (TW，3 行目)．総合不平等を T <- TB + TW で計算し (4 行目)，地域間不平等と地域内不平等の総合不平等に対する寄与度 (contribution) を ContriTB と ContriTW として計算 (5 行目)．以上結果を TDecomp に保存 (5 行目)．

```
group_by(TRegion, year) %>%
    summarize(TB = TheilEntropy(cda / pop, pop),
              TW = sum(w * T),
              T = TB + TW,
              ContriTB = TB / T, ContriTW = TW / T) -> TDecomp
```

この結果が先に計算した THeil の結果と同じであることを確認しておこう．

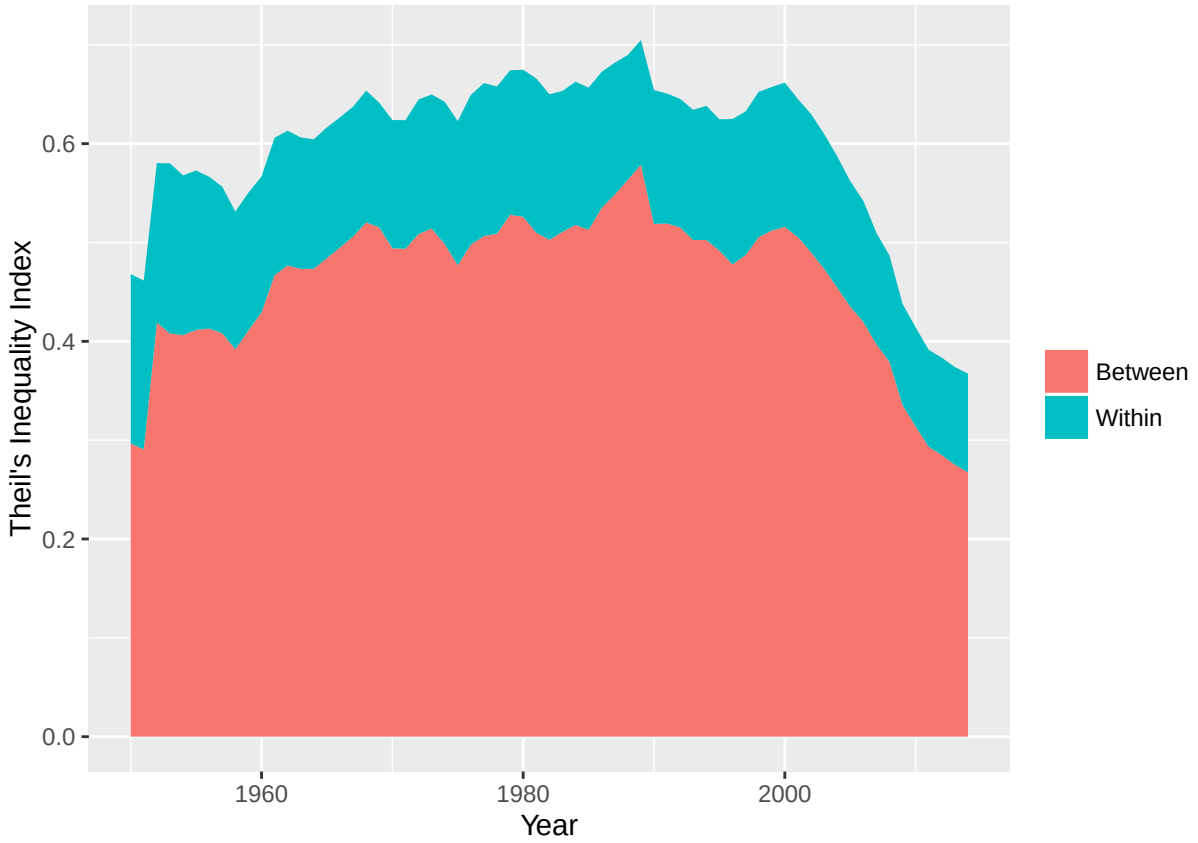
```
all.equal(Theil$T, TDecomp$T)
```

```
## [1] TRUE
```

all.equal は 2 つのオブジェクトが等しいかどうかをテストする関数で，差の絶対値が $1.5 \times 10^{-8}$ より小さければ等しいとみなされる．TRUE が返ってきたので，2 つの計算結果は等しことが確認された．

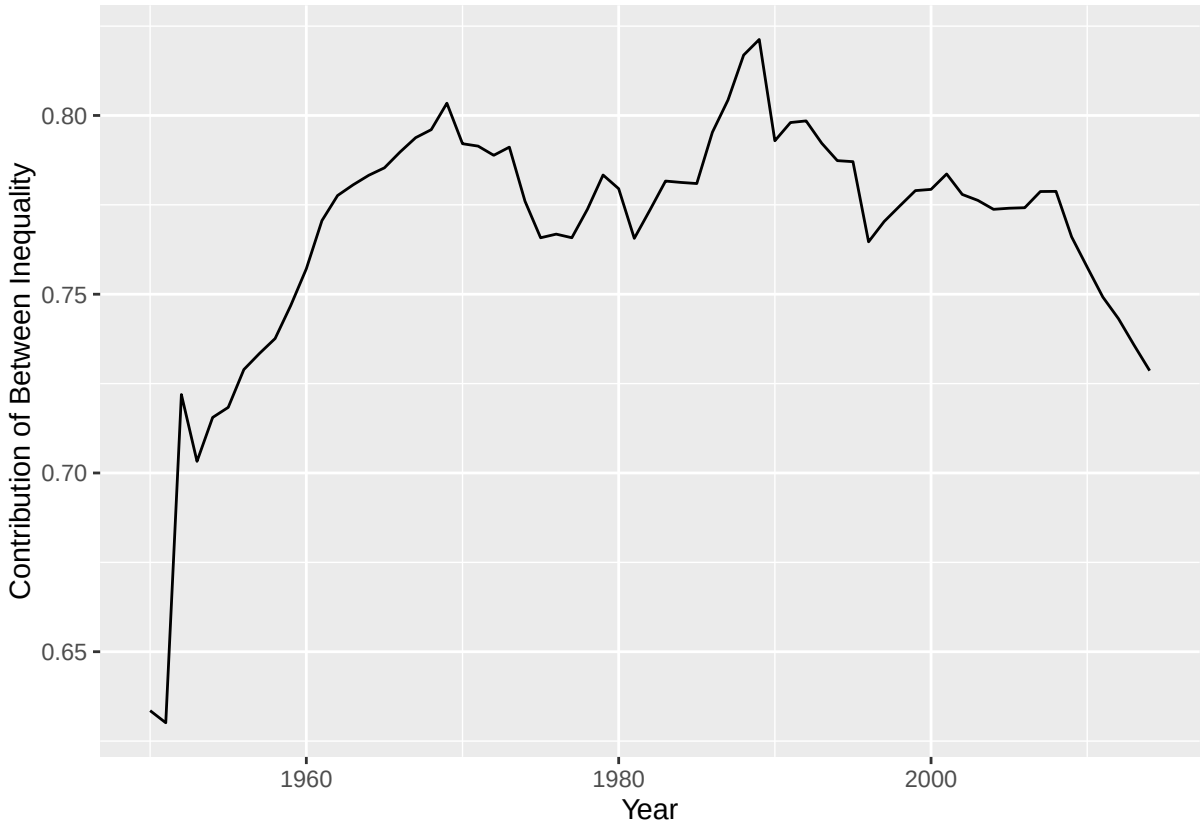
最後に，タイルのエントロピー指数の時系列推移を地域間不平等と地域内不平等の積み上げグラフとして描いてみよう．ggplot に適したロング形式のデータにするために，2 行目で gather を使ってワイド形式からロング形式に変換している．

```
select(TDecomp, year, TB, TW) %>%
gather(Index, T, -year) %>%
ggplot(aes(x = year, y = T, fill = Index)) +
    geom_area() +
    labs(x = "Year", y = "Theil's Inequality Index") +
    scale_fill_discrete(labels=c("Between", "Within")) +
    guides(fill=guide_legend(title = NULL)) -> plotTDecomp
print(plotTDecomp)
```




地域間不平等の寄与度の時系列推移も描いてみよう.

```
ggplot(TDecomp,aes(x = year, y = ContriTB)) +
    geom_line() +
    labs(x = "Year", y = "Contribution of Between Inequality") -> plotContriTB
print(plotContriTB)
```




不平等指数と同様に地域間不平等も 1990 年まで上昇トレンドで，1989 年には 82.1% に至った．それ以降，地域間不平等の寄与度は減少傾向にあるが，2014 年においても地域間不平等が不平等全体の 72.9% を説明している．