## Life as a Pachinko machine: modelling lifetime income using HILDA data

Hugh Parsonage and William Young 2018-07-24

This package outlines the formation of lifetime incomes as used in the retirement income models.

## Outline

The distribution of lifetime income of individuals is calculated by:

- 1. Selecting a relevant subset of the longitudinal sample of the HILDA survey
- 2. For each individual in this subset, calculate their average income between 2005 and 2009 and their average income between 2010 and 2014.
- 3. Determine the income percentile of each individual in each period.
- 4. Construct the observed transition matrix of income for each age transition from 25-29 to 30-34 and from 55-59 to 60-64.
- 5. Smooth these matrix using a tensor product smooth. It may be more appropriate to use a different 2D kernel smoother, but this is the one that was chosen.
- 6. Apply these matrices across each age to end up with a machine\_widget

These are the global variables. That is, only consider HILDA individuals aged between 30 and 70 in the sample and perform 10,000 simulations to obtain our percentiles.

```
# 1000 is pretty good (100,000 individuals)
N_SIMULATIONS <- 1000
OLDEST.AGE <- 70L
YOUNGEST.AGE <- 30L
# tbl_dt [100 x 2]
     Percentile lifetime_AWOTE
#
          \langle int \rangle
                          <dbl>
# 1
                       13.50444
               1
# 2
               2
                       14.17248
# 3
               3
                       14.56721
                       15.00996
               4
# 4
               5
                       15.38401
# 5
# 6
               6
                       15.93115
# 7
              7
                       16.49369
# 8
              8
                       16.91389
# 9
               9
                       17.61472
# 10
             10
                       17.99585
# # ... with 90 more rows
```

```
N_SIMULATIONS <- as.integer(N_SIMULATIONS)</pre>
stopifnot(N_SIMULATIONS >= 1)
randz <- runif(100 * N_SIMULATIONS)</pre>
library(hutils)
library(readr)
library(ggplot2)
library(scales)
library(grattanCharts)
library(viridis)
library(magrittr)
library(mgcv)
library(data.table)
if (requireNamespace("hildaData", quietly = TRUE) &&
    requireNamespace("hildaData2016", quietly = TRUE)) {
  library(hildaData)
  library(hildaData2016)
} else {
  stop("package:hildaData is a local package, containing the raw Combined_* .dta files as data.tables.
       "Either obtain this package from Hugh Parsonage or run", "\n",
       "Combined_n140c <- as.data.table(read.dta('path/to/your/Wave14/Combinedfile.dta'));", "\n",
       "Combined_m130c <- as.data.table(read.dta('path/to/your/Wave13/Combinedfile.dta'));", "\n",
       "\t", "...", "\n",
       "etc.")
library(hildaExtra)
library(grattan)
awote_by_year <-</pre>
  data.table(Year = 2017:2001,
             AWOTE = c(1568,
                        1532,
                        1499,
                        1477,
                        1437,
                        1396,
                        1330.1,
                        1275.2,
                        1226.8,
                        1158,
                        1098.6,
                        1044.1,
                        1011.8,
                        965,
                        929.1,
                        882.1,
                        842.6)) %>%
  .[, AWOTE_annual := 52.5 * AWOTE]
get_hilda_xwaveid_age_income <- function(year){</pre>
```

```
wave <- letters[year - 2000]</pre>
  if (year == 2016L) {
   NAME <- "Combined_p160c"
  } else if (exists("Combined_n150c", where = "package:hildaData")) {
   NAME <- paste0("Combined_", wave, "150c")</pre>
   if (year == 2014){
     NAME <- "Combined n140c"
   } else {
      NAME <- paste0("Combined_", wave, "130c")</pre>
   }
  }
  get(NAME) %>%
   as.data.table %>%
    .[,.SD,
      .SDcols = c("xwaveid",
                  pasteO(wave, "tifefp"), # income
                  paste0(wave, "hgage"), # age
                  pasteO(wave, "esbrd"), # employment-status
                  paste0(wave, "wscmg")
                                          # earnings from wage
      )] %>%
    setnames(paste0(wave, "tifefp"), "Income") %>%
    setnames(pasteO(wave, "hgage"), "Age") %>%
    setnames(paste0(wave, "esbrd"), "Employment_status") %>%
    setnames(paste0(wave, "wscmg"), "Weekly_salary_wage_from_main_job") %>%
    .[, lapply(.SD, make_negatives_NA)] %>%
    .[, Year := get("year", mode = "numeric")]
}
if (exists("Combined_n150c", where = "package:hildaData")) {
  # alias
 Combined_n140c <- as.data.table(Combined_n150c)</pre>
}
xwaveid_by_lnwte <-
 Combined_n140c %>%
  as.data.table %>%
  .[, .(xwaveid, WEIGHT = nlnwte)] %>%
  .[WEIGHT > 0] %>%
  setkey(xwaveid) %>%
  unique(by = "xwaveid")
income_age_by_xwaveid_year <-</pre>
  lapply(2001:2016, get_hilda_xwaveid_age_income) %>%
  rbindlist(use.names = TRUE) %>%
  setkey(xwaveid) %>%
  .[xwaveid by lnwte, on = "xwaveid"] %>%
  .[, Income_percentile := weighted_ntile(Income, WEIGHT, n = 100), keyby = "Year"] %>%
  .[, Income_next_year := shift(Income, type = "lead"), by = "xwaveid"] %>%
  .[, Income_percentile_next_year := shift(Income_percentile, type = "lead"),
  bv = "xwaveid"] %>%
  .[, DOB := Year - Age] %>%
  . []
```

```
prob_year_on_year <-</pre>
  income_age_by_xwaveid_year %>%
  .[between(DOB, 2014L - OLDEST.AGE, 2010L - YOUNGEST.AGE)] %>%
  .[Weekly_salary_wage_from_main_job > 38L] %>%
  .[, Age_group := age_grouper(Age, interval = 10, min_age = 25, max_age = 75)] %>%
  .[, .(n_persons = sum(WEIGHT)),
   keyby = .(Age_group, Income_percentile, Income_percentile_next_year)] %>%
  .[, prob := n_persons / sum(n_persons), keyby = .(Age_group, Income_percentile)] %>%
  . []
Age_group_2005_by_xwaveid <-
  income_age_by_xwaveid_year %>%
  .[Year == 2005L] %>%
  .[, .(Age_group_2005 = age_grouper(Age, interval = 5, min_age = 25, max_age = 75)),
   keyby = "xwaveid"]
## How much do we lose by restricting to only those present in at
## least one year in 2005-2009 and 2010-2014.
ALL <- function(...) {
  Reduce(`&&`, list(...))
}
InScope by xwaveid <-
  income_age_by_xwaveid_year %>%
  .[Weekly_salary_wage_from_main_job > 100L] %>%
  .[Employment_status == "[1] Employed"] %>%
  .[between(DOB, 2014L - OLDEST.AGE, 2010L - YOUNGEST.AGE)] %>%
  .[Year >= 2005L] \%
  .[,
    .(in_scope = ALL(any(Year <= 2009L),</pre>
                     any(Year > 2009L),
                     .N >= 3L)),
   keyby = "xwaveid"]
# Must be at least 60% (of workers at least 30)
stopifnot(mean(InScope_by_xwaveid$in_scope) > 0.6)
income_AWOTE_by_age_percentile <-</pre>
  income_age_by_xwaveid_year %>%
  InScope_by_xwaveid[., on = "xwaveid", nomatch=OL] %>%
  .[(in_scope)] %>%
  .[Year >= 2005L] \%
  awote_by_year[., on = "Year", nomatch=OL] %>%
  .[, Income_per_AWOTE_annual := Income / AWOTE_annual] %>%
    .(AWOTE_annual_average = mean(Income_per_AWOTE_annual)),
   keyby = .(xwaveid, First_year_range = Year <= 2009)] %>%
  setkey(xwaveid) %>%
  xwaveid_by_lnwte[., on = "xwaveid", nomatch=OL] %>%
  Age_group_2005_by_xwaveid[., on = "xwaveid", nomatch=0L] %>%
  .[, AWOTE_annual_percentile := weighted_ntile(AWOTE_annual_average,
                                                 weights = WEIGHT,
```

```
n = 100),
   keyby = .(Age_group_2005, First_year_range)] %>%
  .[, Year_range := if_else(First_year_range, "Decile_ante_2009", "Decile_post_2009"),
   keyby = .(Age_group_2005, First_year_range)] %>%
  .[, .(xwaveid,
        Age_group_2005,
        Year_range,
        AWOTE annual percentile,
        AWOTE_annual_average)] %>%
  setkey(xwaveid) %>%
  xwaveid_by_lnwte[., on = "xwaveid", nomatch=OL] %>%
  .[, .(n_persons = sum(WEIGHT),
        AWOTE_average = weighted.mean(AWOTE_annual_average, WEIGHT)),
   keyby = .(Age_group_2005, Year_range, AWOTE_annual_percentile)]
# income_AWOTE_by_age_percentile %>%
# write_csv("income-AWOTE-by-age-percentile.csv")
prob_5year_on_year <-</pre>
  income_age_by_xwaveid_year %>%
  InScope_by_xwaveid[., on = "xwaveid", nomatch=OL] %>%
  awote_by_year[., on = "Year", nomatch=OL] %>%
  .[Year >= 2005L] %>%
  .[(in scope)] %>%
  .[, Income_per_AWOTE_annual := Income / AWOTE_annual] %>%
  .[,
    .(AWOTE_annual_average = mean(Income_per_AWOTE_annual)),
   keyby = .(xwaveid, First_year_range = Year <= 2009)] %>%
  setkey(xwaveid) %>%
  xwaveid by lnwte[., on = "xwaveid", nomatch=OL] %>%
  Age_group_2005_by_xwaveid[., on = "xwaveid", nomatch=0L] %>%
  .[,
   AWOTE_annual_decile := weighted_ntile(AWOTE_annual_average,
                                          weights = WEIGHT,
                                          n = 10),
   keyby = .(Age_group_2005, First_year_range)] %>%
  .[, Year_range := if_else(First_year_range,
                            "Decile_ante_2009",
                            "Decile_post_2009")] %>%
  .[, .(xwaveid, Age_group_2005, Year_range, AWOTE_annual_decile)] %>%
  dcast.data.table(xwaveid + Age_group_2005 ~ Year_range,
                   value.var = "AWOTE annual decile") %>%
  setkey(xwaveid) %>%
  xwaveid_by_lnwte[., on = "xwaveid", nomatch=0L] %>%
  .[, .(n_persons = sum(WEIGHT)),
   keyby = .(Age group 2005, Decile ante 2009, Decile post 2009)] %>%
  .[, prob := n_persons / sum(n_persons),
   keyby = .(Age_group_2005, Decile_ante_2009)] %>%
  . []
prob_5year_on_year_percentile <-</pre>
  income age by xwaveid year %>%
  InScope_by_xwaveid[., on = "xwaveid", nomatch=OL] %>%
  .[(in_scope)] %>%
```

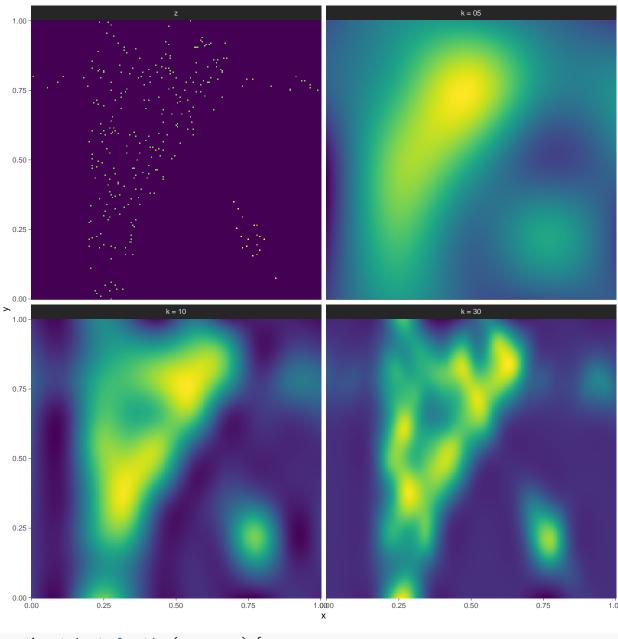
```
.[Year >= 2005L] %>%
  awote_by_year[., on = "Year", nomatch=OL] %>%
  .[, Income_per_AWOTE_annual := Income / AWOTE_annual] %>%
  .[, .(AWOTE_annual_average = mean(Income_per_AWOTE_annual)),
   keyby = .(xwaveid, First_year_range = Year <= 2009)] %>%
  setkey(xwaveid) %>%
  merge(xwaveid_by_lnwte) %>%
  merge(Age group 2005 by xwaveid) %>%
  .[, Year_range := if_else(First_year_range, "Percentile_ante_2009", "Percentile_post_2009")] %>%
  .[, AWOTE_annual_percentile := weighted_ntile(AWOTE_annual_average,
                                                 weights = WEIGHT,
                                                 n = 100),
   keyby = .(Age_group_2005, Year_range)] %>%
  .[, .(xwaveid, Age_group_2005, Year_range, AWOTE_annual_percentile)] %>%
  dcast.data.table(xwaveid + Age_group_2005 ~ Year_range,
                   value.var = "AWOTE_annual_percentile") %>%
  setkey(xwaveid) %>%
  xwaveid_by_lnwte[., on = "xwaveid", nomatch=0L] %>%
  .[, .(n_persons = sum(WEIGHT)),
   keyby = .(Age_group_2005,
              Percentile_ante_2009 = factor(Percentile_ante_2009),
              Percentile_post_2009 = factor(Percentile_post_2009))] %>%
  .[, prob := n_persons / sum(n_persons),
   keyby = .(Age_group_2005, Percentile_ante_2009)] %>%
provide.dir("age-transition-matrices")
nan_to_zero <- function(x) {</pre>
  if (is.numeric(x)){
   x[is.nan(x)] \leftarrow 0
 }
}
get_transition_table <- function(age_range, filename){</pre>
  if (missing(age_range)) {
    age.range \leftarrow gsub("^.*([0-9]\{2\}.[0-9]\{2\}).transition.matrix\.csv$", "\\1", filename)
   fread(filename, header = TRUE) %>%
      melt.data.table(id.vars = "Percentile_pre_2009",
                      variable.name = "Percentile_post_2009",
                      value.name = "prob") %>%
      arrange(Percentile pre 2009) %>%
      group_by(Percentile_pre_2009) %>%
      mutate(cumprob = order_by(Percentile_post_2009, cumsum(prob))) %>%
      setkey(Percentile_pre_2009, cumprob) %>%
      mutate(age_range = age.range)
  } else {
   prob_5year_on_year_percentile[Age_group_2005 == age_range] %>%
      .[, Percentile_ante_2009 := factor(Percentile_ante_2009, levels = 1:100)] %>%
      .[, Percentile_post_2009 := factor(Percentile_post_2009, levels = 1:100)] %>%
      # dcast + melt = expand.grid (so that zero probabilities )
      dcast.data.table(Percentile_ante_2009 ~ Percentile_post_2009,
```

```
fun = mean,
                       drop = FALSE,
                       fill = NaN,
                       value.var = "prob") %>%
      .[, lapply(.SD, nan_to_zero)] %>%
      melt.data.table(id.vars = "Percentile_ante_2009",
                      variable.name = "Percentile_post_2009",
                      value.name = "prob") %>%
      .[, Percentile ante 2009 := as.integer(Percentile ante 2009)] %>%
      .[, Percentile_post_2009 := as.integer(Percentile_post_2009)] %>%
      setkey(Percentile_ante_2009) %>%
      .[, cumprob := cumsum(prob), keyby = "Percentile_ante_2009"] %>%
      setkey(Percentile_ante_2009, cumprob) %>%
      . []
  }
}
transition_tables_all <-</pre>
  lapply(unique(income_AWOTE_by_age_percentile$Age), get_transition_table) %>%
  rbindlist(use.names = TRUE, fill = TRUE)
```

## Tensor product smooth

A tensor product smooth was used to convert the spikes in the sparse transition matrices to a smoothed version, to reflect the underlying distribution. Intuitively, each nonzero value in observed transition matrix is a spike on a square board. The tensor product smooth is the surface that would appear if a somewhat segmented sheet were loosely placed over the spikes on the board. Higher values of k mean the sheet is 'tighter'.

An example of a the tensor product smooth applies to some dummy data is in the following figure. For the model, we used k = 10.



```
smooth_matrix <- function(age_range) {
   stopifnot(length(age_range) == 1L)

gam.1 <-
      prob_5year_on_year_percentile %>%
      .[Age_group_2005 == age_range]

# k = 10 for less smooth data
gam(prob ~ te(Percentile_ante_2009, Percentile_post_2009), k = c(10, 10), data = .)
out <- matrix(pmax(fitted(gam.1), 0), ncol = 100)
out / rowSums(out)
}

smooth_table <- function(age_range) {
   gam.1 <-</pre>
```

```
get_transition_table(age_range) %>%
    # weight2rows("n_persons") %>%
    # k = 10 for less smooth data, somewhat computationally intensive than lower k
    # This is a tensor product smooth.
    gam(prob ~ te(Percentile_ante_2009, Percentile_post_2009, k = c(10, 10)),
        data = .)
  CJ(Percentile ante 2009 = 1:100,
    Percentile_post_2009 = 1:100) %>%
    # gam will predict some values as (very slightly) negative.
    # We need to normalize them anyhow, so just force nonnegative.
    .[, prob := pmax(predict(gam.1, newdata = .), 0)] %>%
    .[, prob := prob / sum(prob), by = Percentile_ante_2009] %>%
    .[, cumprob := cumsum(shift(prob, n = 1, type = "lag", fill = 0)),
     keyby = Percentile_ante_2009] %>%
    .[, .(Percentile_ante = Percentile_ante_2009,
          Percentile_post = Percentile_post_2009,
          cumprob)] %>%
    setkey(Percentile_ante, cumprob)
input <-
  CJ(Percentile_ante = 1:100,
     simulation = 1:N_SIMULATIONS) %>%
  .[, cumprob := randz]
setkey(input, Percentile_ante, cumprob)
copy_column <- function(.data, new_name, column_copied){</pre>
  .data2 <- copy(.data)</pre>
  .data2[[new_name]] <- .data[[column_copied]]</pre>
  as.data.table(.data2)
age_the_simulation <- function(input.table, five_year_group = 2){</pre>
  Ages <- c("25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-59", "60-64")
  age_range_fm <- Ages[five_year_group - 1]</pre>
  agerangefm <- sub("-", "", age_range_fm, fixed = TRUE)</pre>
  age_range_to <- Ages[five_year_group]</pre>
  agerangeto <- sub("-", "", age_range_to, fixed = TRUE)</pre>
  smooth_table(age_range_to)[input.table, roll = Inf] %>%
    setnames("Percentile_post", paste0("Percentile_", agerangeto)) %>%
    setnames("Percentile_ante", paste0("Percentile_", agerangefm)) %>%
    copy_column("Percentile_ante", paste0("Percentile_", agerangefm)) %>%
    # re roll the die
    .[, cumprob := runif(.N)] %>%
    setkey(Percentile_ante, cumprob)
}
machine_widget <-</pre>
  if (YOUNGEST.AGE == 30) {
    input %>%
      setkey(Percentile_ante, cumprob) %>%
      age_the_simulation(2) %>%
```

```
age_the_simulation(3) %>%
      age the simulation(4) %>%
      age_the_simulation(5) %>%
      age_the_simulation(6) %>%
      age_the_simulation(7) %>%
      age_the_simulation(8) %>%
      . []
  } else if (YOUNGEST.AGE == 40) {
    input %>%
      setkey(Percentile_ante, cumprob) %>%
      age_the_simulation(4) %>%
      age_the_simulation(5) %>%
      age_the_simulation(6) %>%
      age_the_simulation(7) %>%
      age_the_simulation(8) %>%
      . []
  } else if (YOUNGEST.AGE == 50) {
    input %>%
      setkey(Percentile_ante, cumprob) %>%
      age_the_simulation(6) %>%
      age_the_simulation(7) %>%
      age_the_simulation(8) %>%
  }
awote by age <-
  income_AWOTE_by_age_percentile[Year_range == "Decile_post_2009",
                                  .(Age = Age_group_2005,
                                   Percentile = AWOTE_annual_percentile,
                                   AWOTE_average)] %>%
  setkey(Age, Percentile) %>%
  # Ensure all percentiles are available
  # (Otherwise the benchmark might be non-monotonic.)
  {
    gridExpanded <- CJ(Age = unique(dot$Age),</pre>
                       Percentile = 1:100)
    setkey(gridExpanded, Age, Percentile)
   dot[gridExpanded, on = c("Age", "Percentile")] %>%
      .[, AWOTE_average := if_else(is.na(AWOTE_average) & Percentile %in% c(1, 100),
                                   if_else(Percentile == 100,
                                           max(AWOTE_average, na.rm = TRUE),
                                   AWOTE_average), keyby = "Age"] %>%
      .[, AWOTE_average := zoo::na.approx(AWOTE_average, na.rm = FALSE), keyby = "Age"]
 }
stopifnot(nrow(awote_by_age) == 100 * (OLDEST.AGE %/% 5 - YOUNGEST.AGE %/% 5))
lifetime income benchmark <-
  awote_by_age %>%
  setkey(Age, Percentile) %>%
```

```
.[, lifetime_AWOTE := 5 * lifetime_AWOTE]
lifetime_pachinko_density <-</pre>
  machine_widget %>%
  .[, .SD, .SDcols = c("simulation",
                       switch(as.character(YOUNGEST.AGE),
                              "30" = "Percentile_2529",
                              "40" = "Percentile_3539",
                              "50" = "Percentile_4549",
                              stop("Unanticipated YOUNGEST.AGE")),
                       grep(".1", names(.), value = TRUE),
                       switch(as.character(OLDEST.AGE),
                              "70" = "Percentile_6064",
                              stop("Unanticipated OLDEST.AGE")))] %>%
  setnames(old = grep(".1$", names(.), value = TRUE),
           new = gsub(".1$", "", names(.)[grep(".1$", names(.), value = FALSE)])) %>%
  .[, id := .I] \%
  melt.data.table(id.vars = c("id", "simulation"), variable.name = "Age", value.name = "Percentile") %>
  .[, Age := gsub("Percentile_(..)(..)", "\\1-\\2", Age)] %>%
  setkeyv(c("Age", "Percentile")) %>%
  merge(awote_by_age)
lifetime_income_pachinko <-</pre>
  lifetime_pachinko_density[,
                             .(lifetime_AWOTE = sum(AWOTE_average) * 5,
                              Percentile = first(Percentile)),
                            keyby = "id"] %>%
  .[, .(lifetime_AWOTE = mean(lifetime_AWOTE)), keyby = "Percentile"] %>%
  .[, lifetime_AWOTE := round(lifetime_AWOTE, 1)] %T>%
  fwrite(sprintf("../inst/extdata/lifetime_income_pachinko-Age%s.csv", YOUNGEST.AGE)) %>%
 . []
```

.[, .(lifetime\_AWOTE = sum(AWOTE\_average)), keyby = "Percentile"] %>%

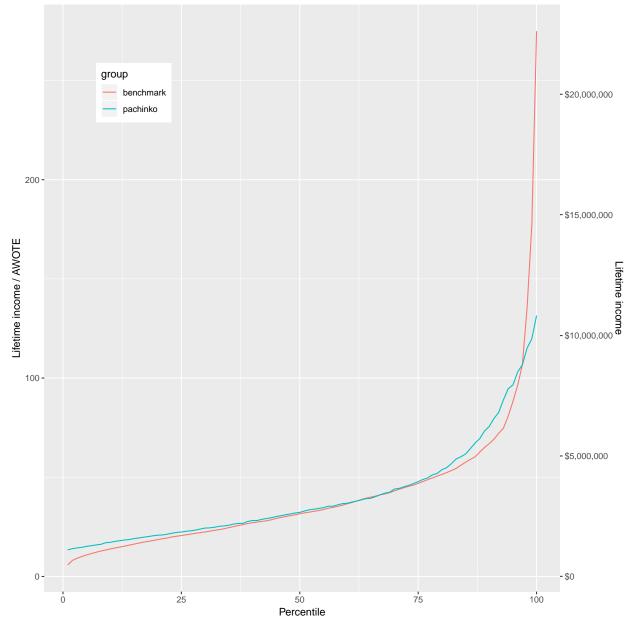
## Result

The average lifetime income is comparable to the average income if we assumed individuals remained in the same percentile throughout their lives.

```
merge(lifetime_income_benchmark, lifetime_income_pachinko, by = "Percentile") %$%
    t.test(lifetime_AWOTE.x, lifetime_AWOTE.y)

#>
#> Welch Two Sample t-test
#>
#> data: lifetime_AWOTE.x and lifetime_AWOTE.y
#> t = -0.069473, df = 180.13, p-value = 0.9447
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#> -8.911115 8.304976
#> sample estimates:
#> mean of x mean of y
#> 40.06493 40.36800
```

The difference is most pronounced at the tails of the distribution:



```
for (the_percentile in c(10, 25, 50, 95)) {
   the_percentile_plot <-</pre>
```

```
switch(as.character(YOUNGEST.AGE),
         "30" = {
           machine_widget[Percentile_2529 %in% the_percentile,
                          .SD,
                          .SDcols = grep("^((Percentile_[0-9]{4})))(.*(6064|2529)))",
                                         names(machine_widget)),
                          key = "simulation"]
         },
         "40" = {
          machine_widget[Percentile_3539 %in% the_percentile,
                          .SDcols = grep("^((Percentile_[0-9]{4})\.1)|(.*(6064|3539)))",
                                         names(machine_widget)),
                          key = "simulation"]
         },
         "50" = {
           machine_widget[Percentile_4549 %in% the_percentile,
                          .SDcols = grep("^((Percentile_[0-9]{4})))(.*(6064|4549)))",
                                         names(machine_widget)),
                          key = "simulation"]
         }) %>%
  melt.data.table(id.vars = "simulation") %>%
  .[, variable := sub("^Percentile_(..)(..).*$", "\\1-\\2", variable)] %>%
  .[order(-variable)] %>%
  .[, variable := factor(variable,
                         levels = rev(unique(.$variable)),
                         ordered = TRUE)] %>%
  ggplot(aes(x = value, fill = variable)) +
  geom_bar(width = 1,
           color = grey(level = 0.75),
           aes(y = ..prop..)) +
  facet_wrap( ~ variable, scales = "free_y", ncol = 1) +
  theme_dark() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = percent, breaks = c(0, 0.02, 0.04, 0.06, 0.08))
print(the_percentile_plot)
```

