Analysis for "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers"

Neil Stewart

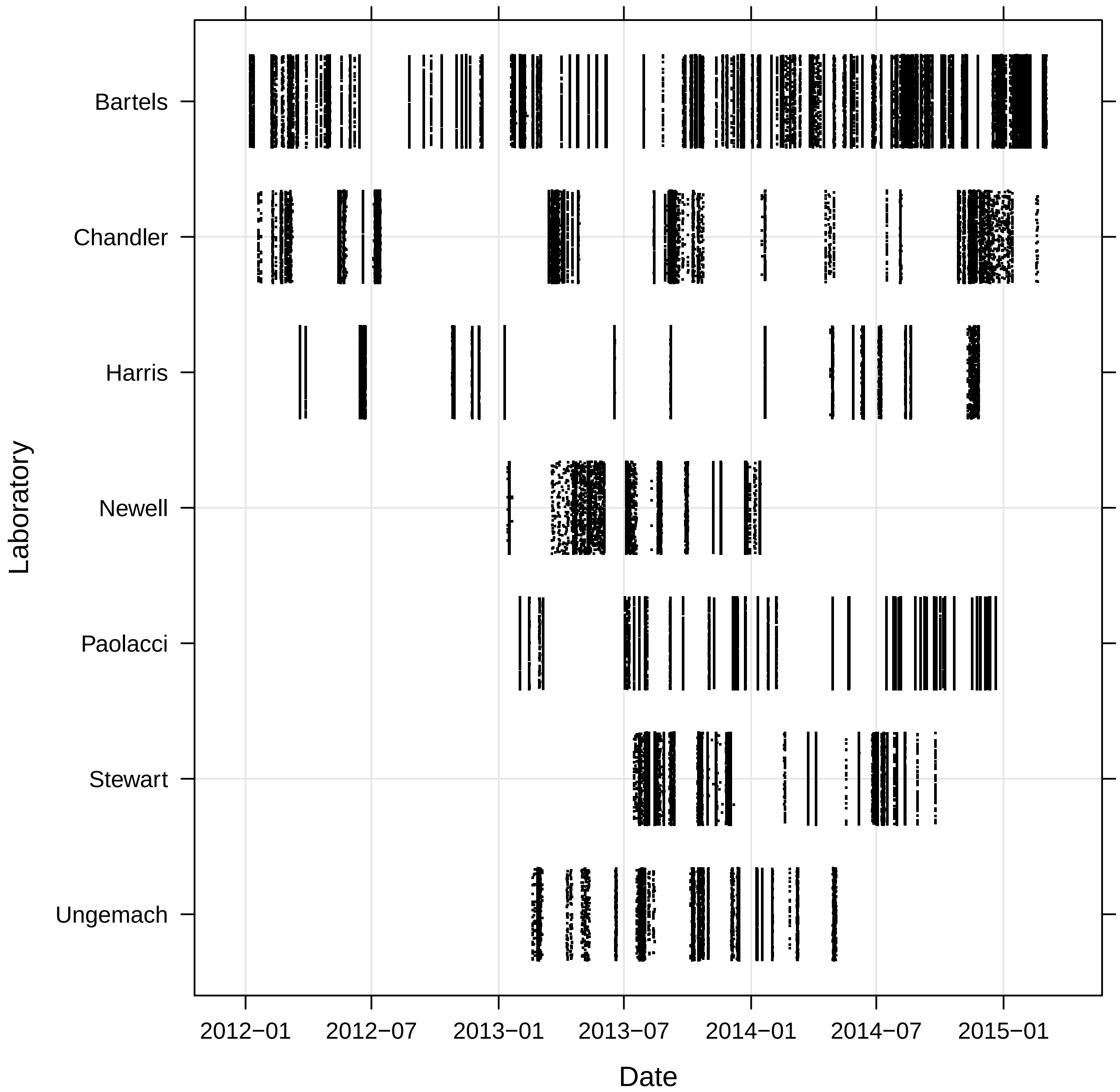
1 Load data

```
# Code for making the web version of HITs.RData
load("../data_2/HITs.RData")
HITs <- HITs[, .(WorkerId, SubmitTime, WorkTimeInSeconds, filename, location.requirement, HIT.requirement, conditional, pay, median.duration,
  lab)]

UUIDs <- data.table(WorkerId = unique(HITs$WorkerId))
UUIDs$UUID <- replicate(n = nrow(UUIDs), UUIDgenerate(use.time = FALSE))

HITs <- merge(HITs, UUIDs, by = "WorkerId")
HITs <- HITs[, `:=`(WorkerId, NULL)]
setnames(HITs, "UUID", "WorkerId")
save(HITs, file = "HITs.RData")
write.csv(HITs, file = "HITs.csv", row.names = FALSE)
```

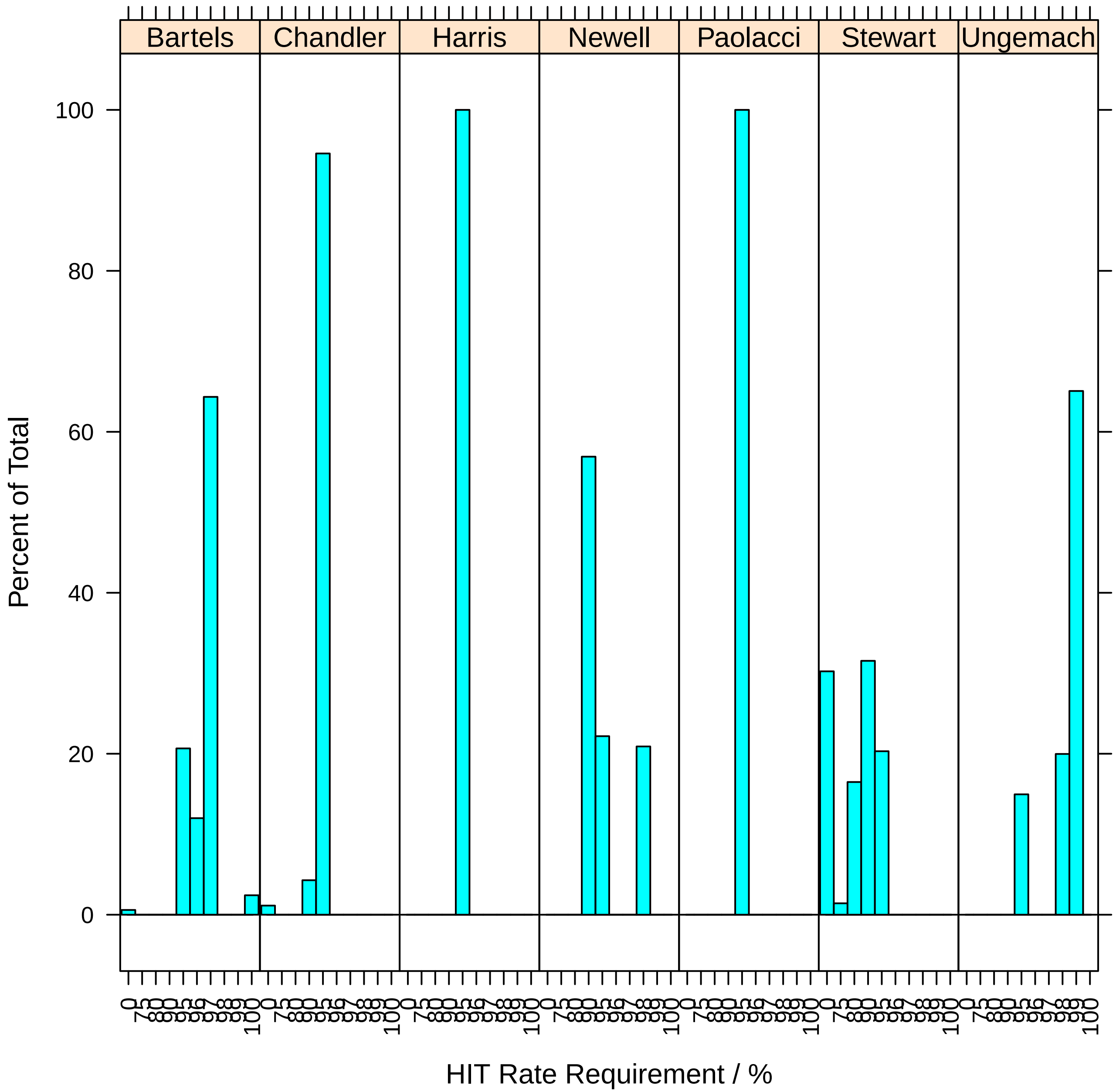
```
load("HITs.RData")

# HITs.RData contains the data.table HITs, with colums: WorkerId --- Each unique WorkerId has been swapped for a UUID SubmitTime --- (Column
# from original MTurk Batch file) WorkTimeInSeconds --- (Column from original MTurk Batch file) filename --- The name of the MTurk Batch file
# location.requirement --- Location requirement for the HIT (self report from the experimenter) HIT.requirement --- HIT approval rate
# requirement (self report from the experimenter) conditional --- Whether the experiment required participation in an earlier study (self
# report from the experimenter) pay --- in dollars, stripped from the Reward column in the original MTurk Batch file median.duration --- The
# median WorkTimeInSeconds for each batch lab --- The surname of the experimenter supplying the data
```

2 Section 2: The Laboratories

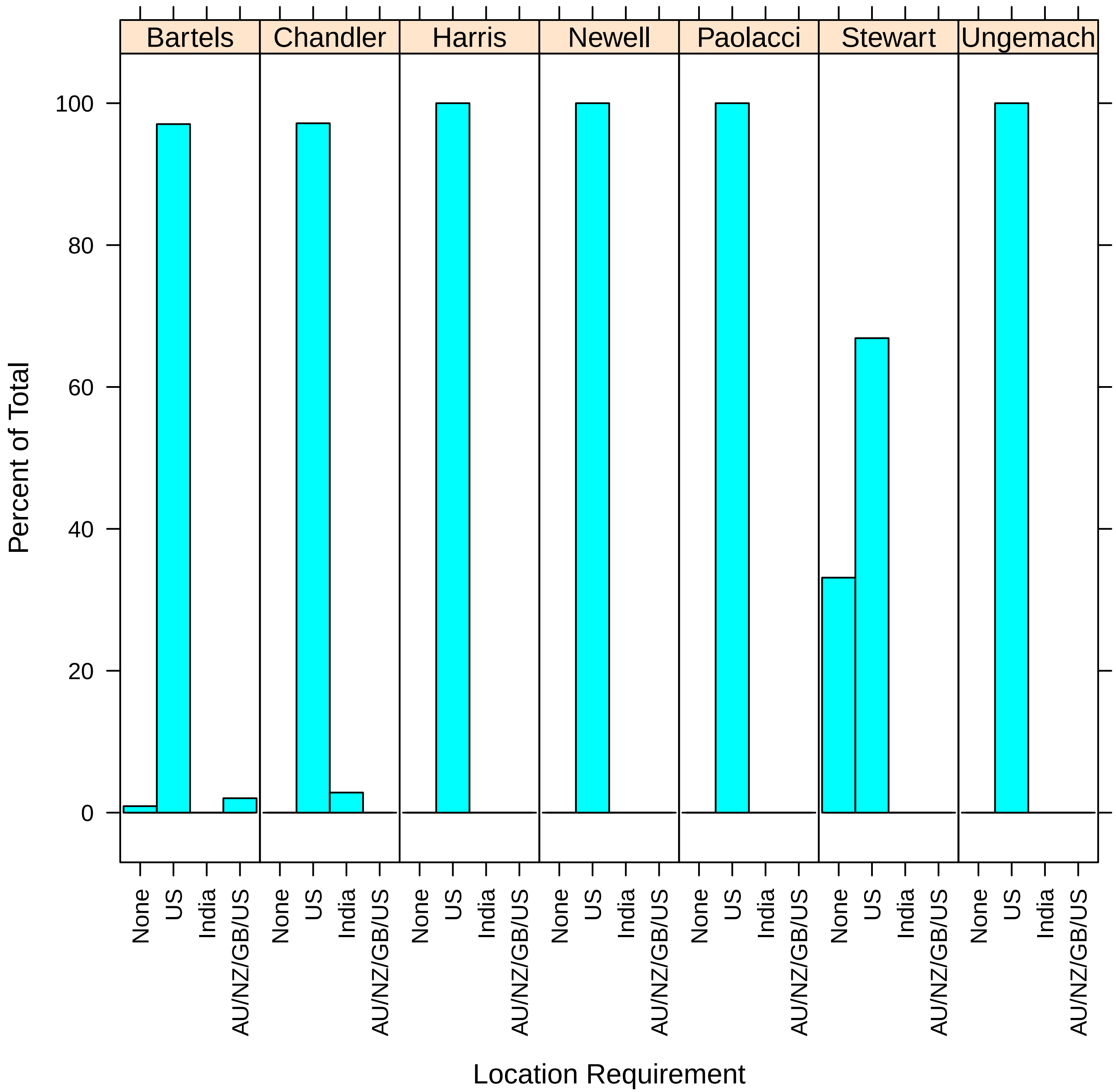
```r
# Number of HITs
nrow(HITs)

## [1] 114460

# Number of unique workers
length(unique(HITs$WorkerId))

## [1] 33408

# Number of batches
length(unique(HITs$filename))

## [1] 689

# Time range of HITs
(time.range <- range(HITs$SubmitTime, na.rm = TRUE))

## [1] "2012-01-07 18:44:11 GMT" "2015-03-03 20:48:05 GMT"

HITs <- HITs[, `:=`(lab.name, as.factor(lab))]
(date.plot <- xyplot(factor(lab.name, levels = rev(levels(lab.name))) ~ SubmitTime, data = HITs, type = c("p", "g"), pch = ".", jitter.y = TRUE,
  xlab = "Date", ylab = "Laboratory", factor = 1.7, col = "black"))
```


Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ungemach
Laboratory
2012−01
2012−07
2013−01
2013−07
2014−01
2014−07
2015−01
Date

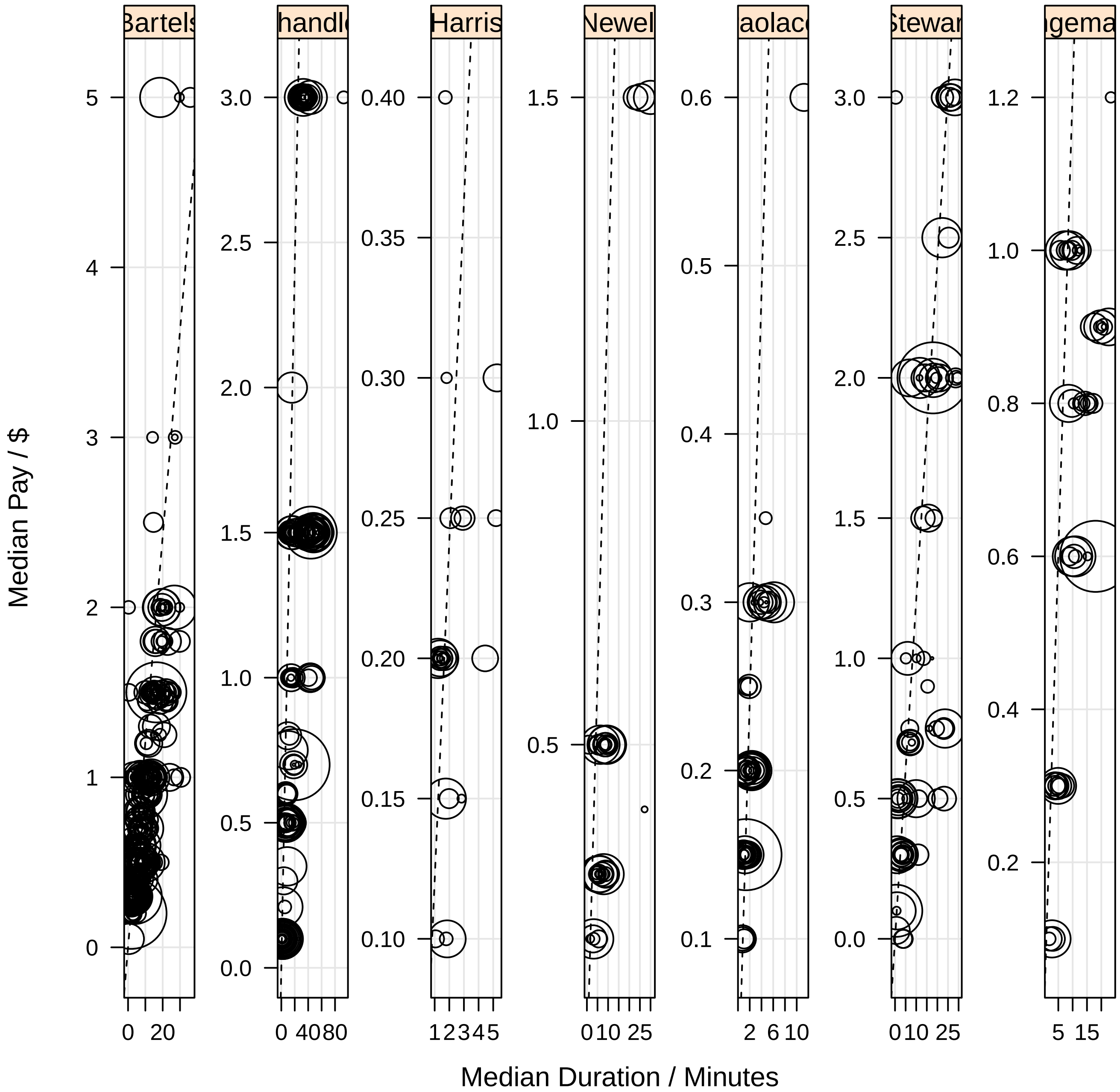
```r
HITs <- HITs[is.na(HIT.requirement), `:=`(HIT.requirement, 0)]
(hit.plot <- histogram(~as.factor(HIT.requirement) | lab.name, data = HITs, layout = c(7, 1), scales = list(alternating = FALSE, x = list(rot = 90)),
  xlab = "HIT Rate Requirement / %"))
```


Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ungemach
Percent of Total
100
80
60
40
20
0
HIT Rate Requirement / %

```
(location.plot <- histogram(~factor(location.requirement, levels = c("", "UNITED STATES", "INDIA", "AU, NZ, GB, UNITED STATES"), labels = c("None",
  "US", "India", "AU/NZ/GB/US")) | lab.name, data = HITs, layout = c(7, 1), scales = list(alternating = FALSE, x = list(rot = 90)), xlab = "Location Require
```


Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ungemach
Percent of Total
0
20
40
60
80
100
None
US
India
AU/NZ/GB/US
Location Requirement

```r
HITs <- HITs[, `:=`(median.duration, median(WorkTimeInSeconds)), by = filename]
HITs <- HITs[, `:=`(median.payrate, pay/median.duration * 60 * 60)]  # Dollars per hour
median.pay.by.duration <- HITs[, .(median.duration = median(WorkTimeInSeconds), N = .N, pay = median(pay)), by = .(filename, lab.name)]

# Median duration in minutes
median(HITs$WorkTimeInSeconds)/60

## [1] 4.416667

# Median pay
median(HITs$pay)

## [1] 0.35

# Median hourly pay
median(with(HITs, pay/WorkTimeInSeconds * 60 * 60))

## [1] 5.538462

(pay.by.duration.plot <- xyplot(pay ~ median.duration/60 | lab.name, data = median.pay.by.duration, cex = sqrt(median.pay.by.duration$N)/10,
  scales = list(alternating = FALSE, relation = "free", y = list(rot = 0)), layout = c(7, 1), xlab = "Median Duration / Minutes", ylab = "Median Pay / $",
  col = "black", type = c("p", "g")) + layer(panel.abline(a = 0, b = 7.25/60, lty = 2)))
```


Bartels
handl
Harris
Newel
aolac
Stewar
gema
Median Pay / $
Median Duration / Minutes

```r
# Figure 1
h <- 0.2
pdf("lab_details_2.pdf", height = 14, width = 14 * 210/297 * 1.2)
print(date.plot, position = c(0, 3 * h, 1, 1))
print(hit.plot, position = c(0, 2 * h, 1, 3 * h), newpage = FALSE)
print(location.plot, position = c(0, 1 * h, 1, 2 * h), newpage = FALSE)
print(pay.by.duration.plot, position = c(0, 0 * h, 1, 1 * h), newpage = FALSE)
dev.off()
```

```
## pdf
##   2
```

3 Section 3: The Size of the MTurk Population

```r
# Back up HITs data.table for later sections
HITs.original <- HITs

# Add which year quarter
HITs <- HITs[, `:=`(quarter, cut(SubmitTime, "quarter"))]

# data.table of workers in each
workers.per.batch <- HITs[, .(no.workers = length(unique(WorkerId)), no.HITs = .N), by = .(lab, filename)]
workers.per.batch <- workers.per.batch[, `:=`(HITs.per.worker, no.HITs/no.workers)]

# Batches allowing multiple submissions
multiple.response.filenames <- workers.per.batch[HITs.per.worker > 1.1]$filename
HITs <- HITs[, `:=`(multiple.responses, ifelse(filename %in% multiple.response.filenames, "Yes", "No"))]

########################################################## Restrict to open experiments without multiple.responses
HITs <- HITs[conditional == "Open" & multiple.responses == "No"]
nrow(HITs)/nrow(HITs.original)
```

```
## [1] 0.8097152
```

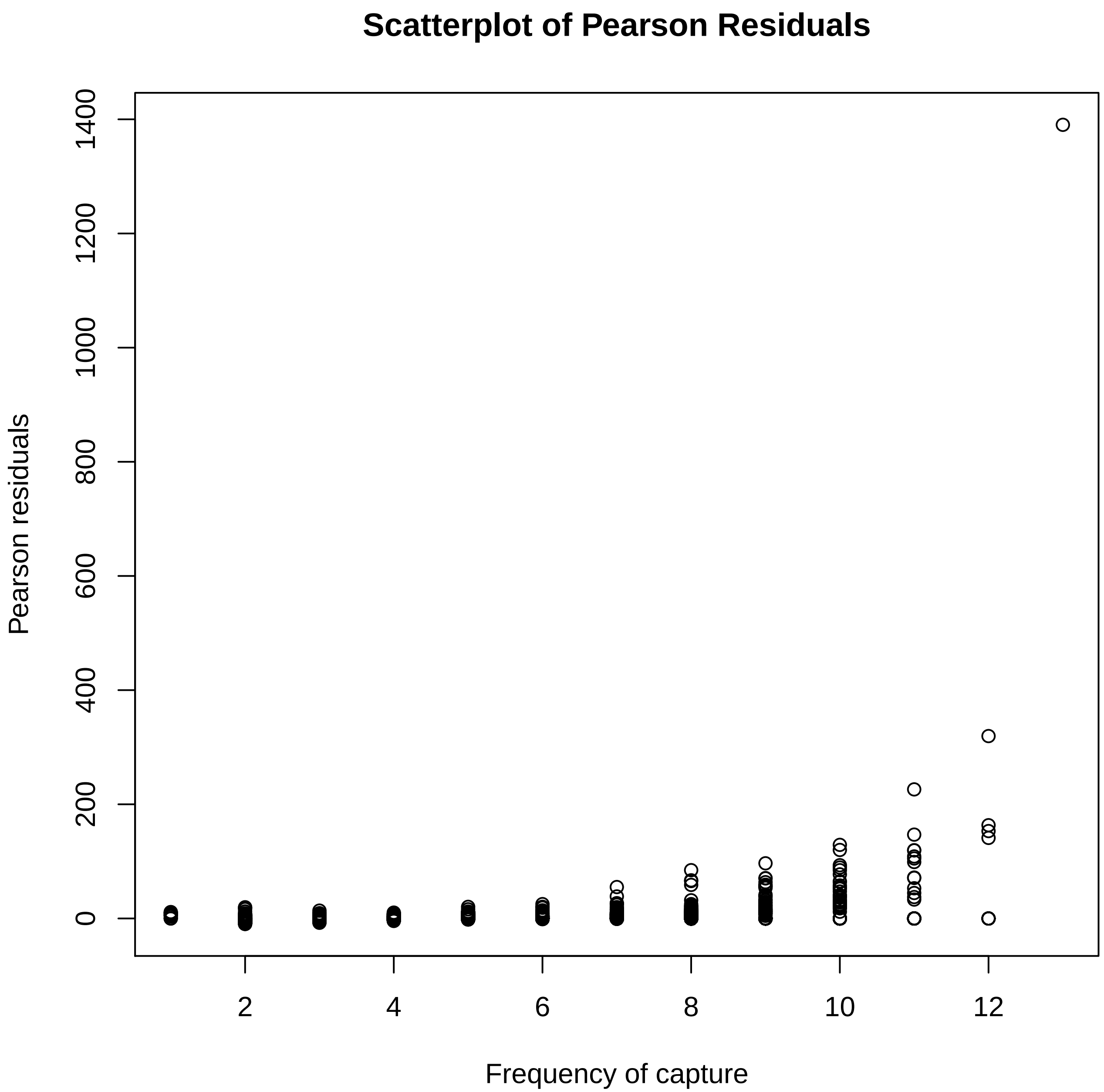
```r
# The All-Labs estimate

cap.recap.openp <- function(HITs, lab = NA, ...) {
  # Wrapper to run open-population analysis with descriptive(), capture histories, and openp() HITs is a data.frame with one row per capture,
  # with columns for WorkerId and quarter
  capture.histories <- xtabs(~WorkerId + quarter, data = HITs)
  capture.histories[capture.histories > 1] <- 1
  capture.histories <- capture.histories[, colSums(capture.histories) > 0]  # Delete columns for occasions when no one is caught
```

```
  results <- list(periods = colnames(capture.histories))
  results$descriptive = descriptive(capture.histories)
  if (lab == "get.from.HITs.data.table")
    results$lab <- HITs$lab[1] else results$lab <- lab
  if (ncol(capture.histories) > 3) {
    results$openp <- openp(capture.histories, ...)
    # if(!missing(keep)) results£capture.history.freqs <- cbind(histpos.t(ncol(capture.histories)), results£openp£glm£model£Y)
  }
  return(results)
}

# Run the open-population analysis on data from all laboratories
(op.all <- cap.recap.openp(HITs, lab = "All Labs"))

## $periods
##  [1] "2012-01-01" "2012-04-01" "2012-07-01" "2012-10-01" "2013-01-01" "2013-04-01" "2013-07-01" "2013-10-01"
##  [9] "2014-01-01" "2014-04-01" "2014-07-01" "2014-10-01" "2015-01-01"
##
## $descriptive
##
## Number of captured units: 31013
##
## Frequency statistics:
##          fi    ui    vi    ni
## i = 1   21180  1980  1239  1980
## i = 2    5259  2869  2423  3326
## i = 3    2026   277   128   378
## i = 4     992  2009  1354  2518
## i = 5     610  1869  1361  2718
## i = 6     382  2422  1647  3493
## i = 7     256  3976  2982  5867
## i = 8     129  3909  3861  6631
## i = 9      83  1271  1565  3224
## i = 10     57  2501  1928  4353
## i = 11     26  3155  3720  6350
## i = 12      7  3287  4695  6727
## i = 13      6  1488  4110  4110
## fi: number of units captured i times
## ui: number of units captured for the first time on occasion i
## vi: number of units captured for the last time on occasion i
## ni: number of units captured on occasion i
##
##
## $lab
```

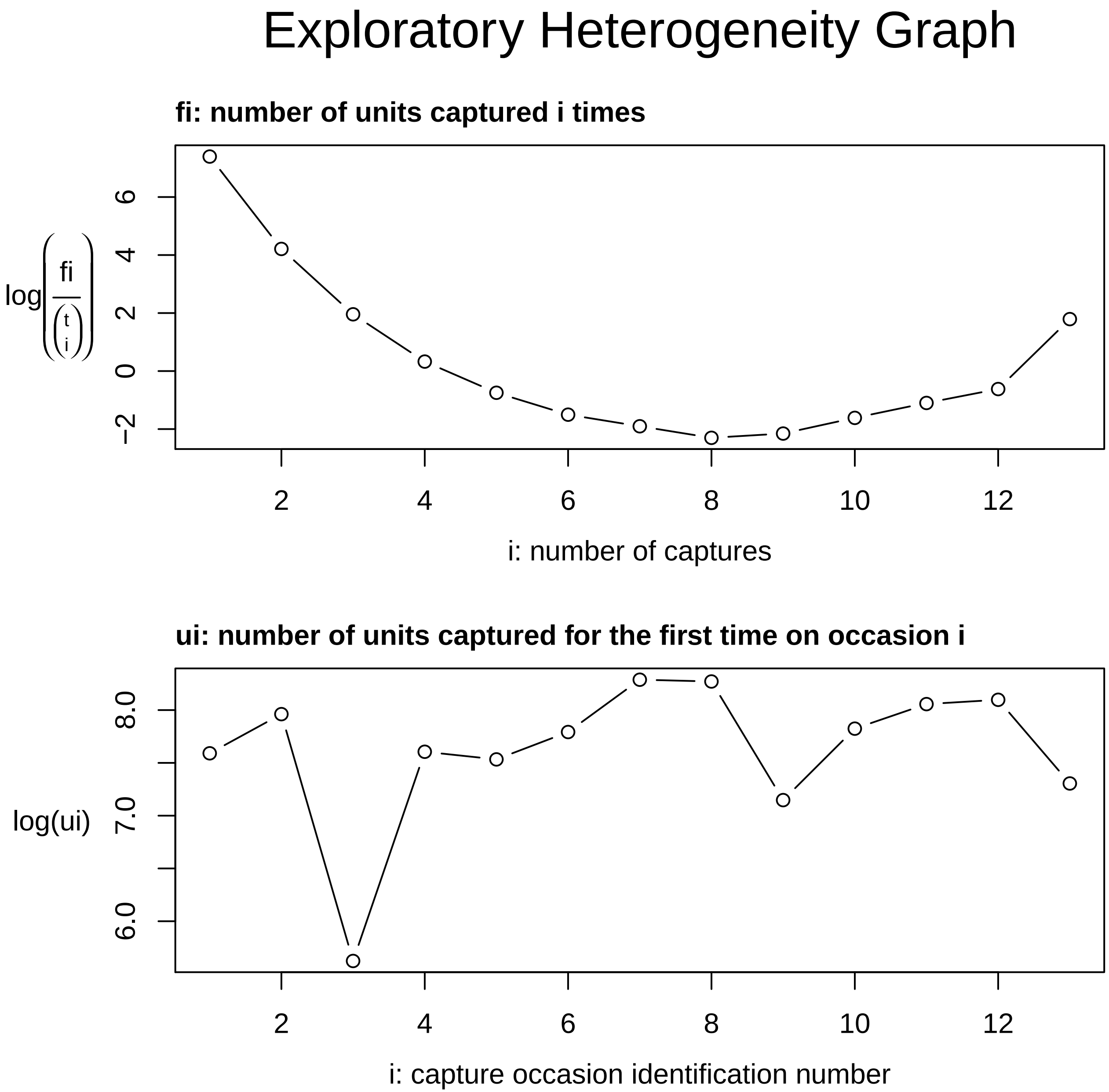
```
## [1] "All Labs"
##
## $openp
##
## Model fit:
##               deviance     df        AIC
## fitted model  11437.74    8156   14911.85
##
## Test for trap effect:
##                                    deviance     df        AIC
## model with homogenous trap effect   9342.96    8155   12819.07
## model with trap effect              9227.95    8146   12722.05
##
## Capture probabilities:
##            estimate  stderr
## period 1         --      --
## period 2     0.3040  0.0171
## period 3     0.0524  0.0052
## period 4     0.2307  0.0102
## period 5     0.2707  0.0097
## period 6     0.2839  0.0088
## period 7     0.4020  0.0091
## period 8     0.4238  0.0085
## period 9     0.2984  0.0073
## period 10    0.3222  0.0075
## period 11    0.4859  0.0088
## period 12    0.6378  0.0111
## period 13        --      --
##
## Survival probabilities:
##                  estimate  stderr
## period 1 -> 2      0.7591  0.0321
## period 2 -> 3      0.4412  0.0145
## period 3 -> 4      1.0000  0.0000
## period 4 -> 5      0.7441  0.0188
## period 5 -> 6      0.7538  0.0174
## period 6 -> 7      0.7592  0.0143
## period 7 -> 8      0.7400  0.0127
## period 8 -> 9      0.6335  0.0123
## period 9 -> 10     0.7353  0.0146
## period 10 -> 11    0.7972  0.0135
## period 11 -> 12    0.5543  0.0105
## period 12 -> 13        --      --
```

```
##
## Abundances:
##            estimate  stderr
## period 1        --      --
## period 2   10939.1   593.8
## period 3    7219.2   623.8
## period 4   10912.6   444.6
## period 5   10040.8   320.7
## period 6   12305.6   340.1
## period 7   14593.0   295.0
## period 8   15646.7   279.7
## period 9   10804.6   212.5
## period 10  13508.5   266.1
## period 11  13068.1   206.4
## period 12  10546.6   167.3
## period 13       --      --
##
## Number of new arrivals:
##                  estimate  stderr
## period 1 -> 2         --      --
## period 2 -> 3     2392.9   644.2
## period 3 -> 4     3693.5   713.2
## period 4 -> 5     1920.6   390.7
## period 5 -> 6     4736.8   347.6
## period 6 -> 7     5250.2   320.8
## period 7 -> 8     4847.7   267.7
## period 8 -> 9      892.7   185.7
## period 9 -> 10    5563.8   244.1
## period 10 -> 11   2299.6   211.0
## period 11 -> 12   3303.2   132.1
## period 12 -> 13       --      --
##
## Total number of units who ever inhabited the survey area:
##              estimate  stderr
## all periods   47241.3   410.7
##
## Total number of captured units: 31013
```

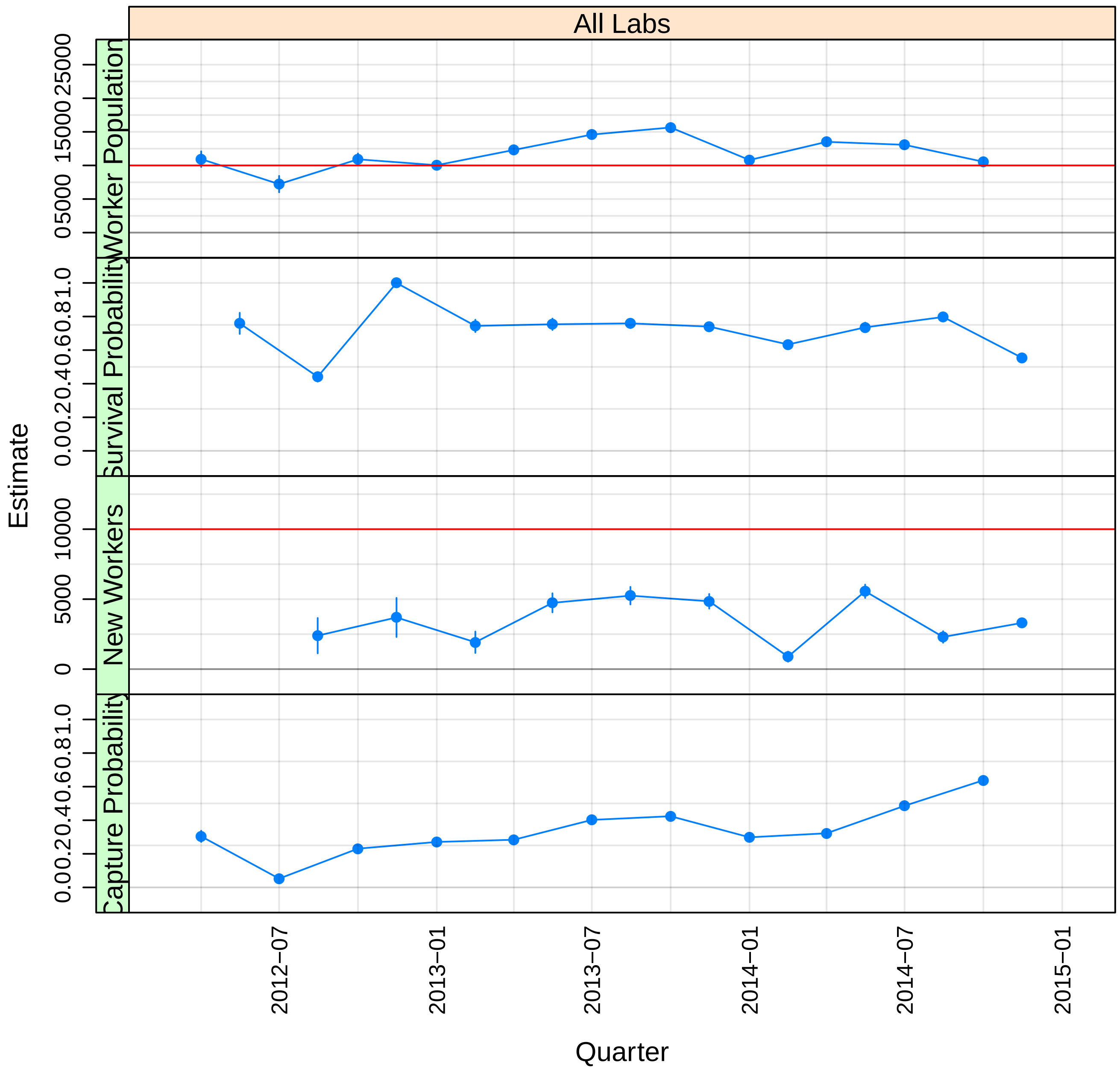
```
plot(op.all$openp)
```


Scatterplot of Pearson Residuals
Pearson residuals
0
200
400
600
800
1000
1200
1400
2
4
6
8
10
12
Frequency of capture

```
plot(op.all$descriptive)
```


Exploratory Heterogeneity Graph
fi: number of units captured i times
log(fi / (t choose i))
−2
0
2
4
6
2
4
6
8
10
12
i: number of captures
ui: number of units captured for the first time on occasion i
log(ui)
6.0
7.0
8.0
2
4
6
8
10
12
i: capture occasion identification number

```r
openp.df <- function(op) {
  # Convert openp() output to data.frame with confidence intervals
  add.CIs <- function(d, type, periods) {
    d <- as.data.frame(d)
    d$type <- type
    d$lower <- with(d, estimate - qnorm(0.975) * stderr)
    d$upper <- with(d, estimate + qnorm(0.975) * stderr)
    if (type %in% c("Survival Probability", "New Workers")) {
      d$period <- periods[-1]
      d$period <- as.POSIXct(d$period, "%Y-%m-%d")
      d$period <- d$period + 3600 * 24 * 45
    } else {
      d$period <- periods
      d$period <- as.POSIXct(d$period, "%Y-%m-%d")
    }
    d
  }
  capture.probs <- add.CIs(op$openp$capture.prob, "Capture Probability", op$periods)
  survival.probs <- add.CIs(op$openp$survivals, "Survival Probability", op$periods)
  new.arrivals <- add.CIs(op$openp$birth, "New Workers", op$periods)
  abundance <- add.CIs(op$openp$N, "Worker Population", op$periods)

  d <- rbind(capture.probs, survival.probs, new.arrivals, abundance)
  d$lab <- op$lab
  d
}

qs <- as.POSIXct(unique(HITs$quarter), "%Y-%m-%d")
combineLimits(useOuterStrips(segplot(period ~ lower + upper | "All Labs" + type, centers = estimate, data = openp.df(op.all), horizontal = FALSE,
  xlab = "Quarter", ylab = "Estimate", scales = list(y = list(relation = "free"), x = list(rot = 90), alternating = FALSE), type = "b", ylim = list(c(0,
    1), c(0, 12000), c(0, 1), c(0, 25000)), xlim = time.range) + layer(panel.abline(h = c(seq(0, 1, 0.25), seq(0, 25000, 2500)), alpha = 0.1))) +
  layer(panel.abline(v = qs, alpha = 0.1)) + layer(panel.abline(h = 10000, col = "red")))
```



All Labs
Worker Population
Survival Probability
New Workers
Capture Probability
Estimate
0
5000
15000
25000
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
2012−07
2013−01
2013−07
2014−01
2014−07
2015−01
Quarter

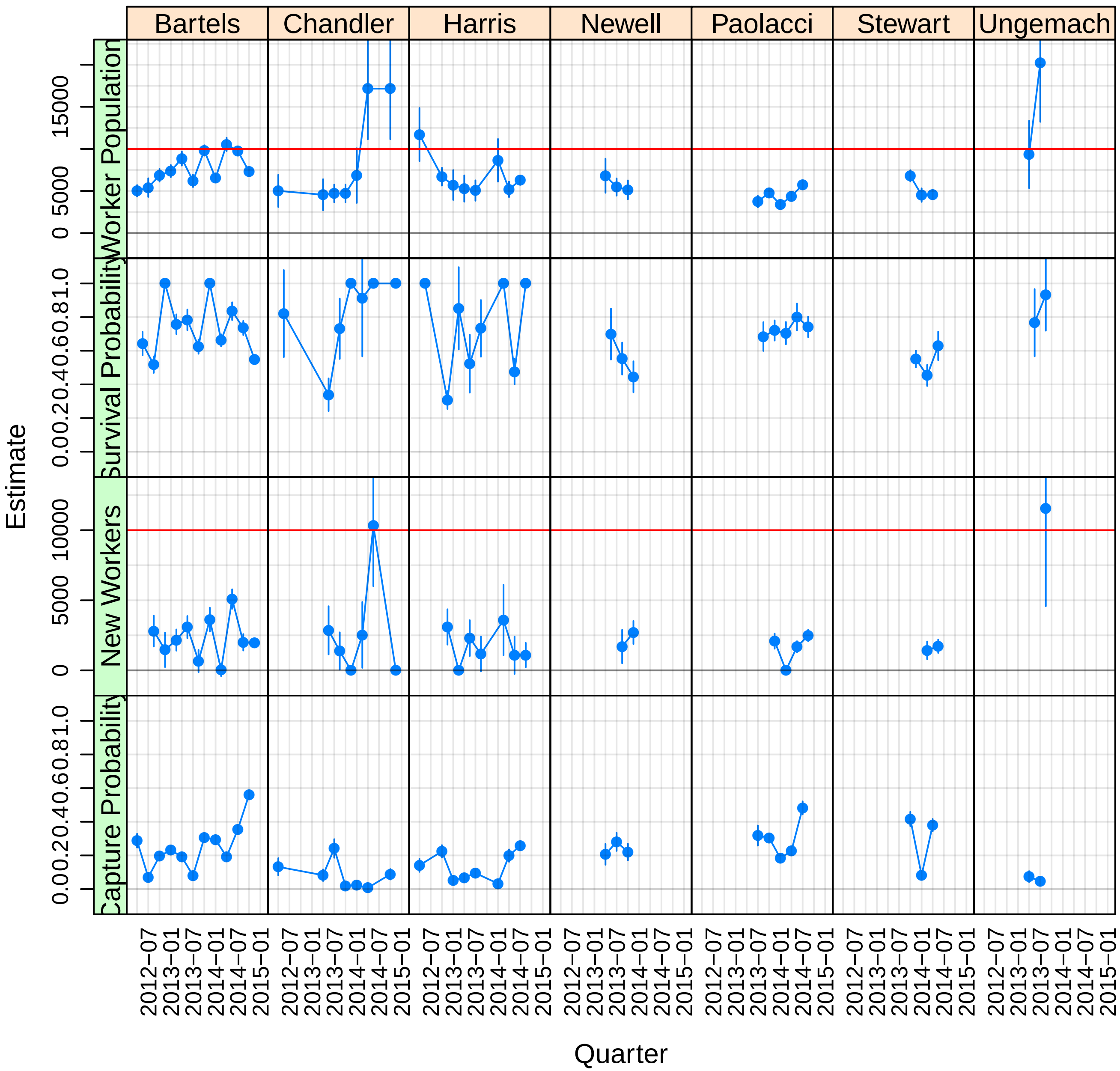
```
# Mean population size estimate over periods
mean(op.all$openp$N[, "estimate"], na.rm = TRUE)

## [1] 11780.44
```

```
# Separate estimates for each lab

# Use by() to run the open-population analysis separately for each lab
op <- by(HITs, INDICES = list(HITs$lab), FUN = cap.recap.openp, lab = "get.from.HITs.data.table")
# Print results for one lab
op$Bartels

## $periods
##  [1] "2012-01-01" "2012-04-01" "2012-07-01" "2012-10-01" "2013-01-01" "2013-04-01" "2013-07-01" "2013-10-01"
##  [9] "2014-01-01" "2014-04-01" "2014-07-01" "2014-10-01" "2015-01-01"
##
## $descriptive
##
## Number of captured units: 17633
##
## Frequency statistics:
##          fi     ui     vi     ni
## i = 1   12611   1495   1022   1495
## i = 2    2932   1166   1027   1442
## i = 3    1005    301    174    378
## i = 4     494   1069    803   1343
## i = 5     299   1277   1033   1712
## i = 6     133   1220   1032   1691
## i = 7      71    313    200    498
## i = 8      35   2196   1762   2993
## i = 9      29    981   1021   1915
## i = 10     13   1354   1103   2021
## i = 11      8   2198   2049   3458
## i = 12      3   2330   2403   4101
## i = 13      0   1733   4004   4004
## fi: number of units captured i times
## ui: number of units captured for the first time on occasion i
## vi: number of units captured for the last time on occasion i
## ni: number of units captured on occasion i
##
##
## $lab
## [1] "Bartels"
##
## $openp
```

```
##
## Model fit:
##              deviance     df        AIC
## fitted model  5465.875   8157   7726.076
##
## Test for trap effect:
##                                    deviance     df        AIC
## model with homogenous trap effect  4960.335   8156   7222.537
## model with trap effect             4942.148   8147   7222.349
##
## Capture probabilities:
##            estimate  stderr
## period 1         --      --
## period 2     0.2873  0.0209
## period 3     0.0700  0.0081
## period 4     0.1955  0.0120
## period 5     0.2324  0.0125
## period 6     0.1909  0.0102
## period 7     0.0804  0.0060
## period 8     0.3048  0.0109
## period 9     0.2934  0.0106
## period 10    0.1918  0.0083
## period 11    0.3545  0.0109
## period 12    0.5613  0.0127
## period 13        --      --
##
## Survival probabilities:
##                  estimate  stderr
## period 1 -> 2      0.6425  0.0357
## period 2 -> 3      0.5174  0.0249
## period 3 -> 4      1.0000  0.0000
## period 4 -> 5      0.7579  0.0298
## period 5 -> 6      0.7835  0.0315
## period 6 -> 7      0.6242  0.0214
## period 7 -> 8      1.0000  0.0000
## period 8 -> 9      0.6617  0.0184
## period 9 -> 10     0.8353  0.0269
## period 10 -> 11    0.7356  0.0216
## period 11 -> 12    0.5484  0.0131
## period 12 -> 13        --      --
##
## Abundances:
##            estimate  stderr
## period 1         --      --
```

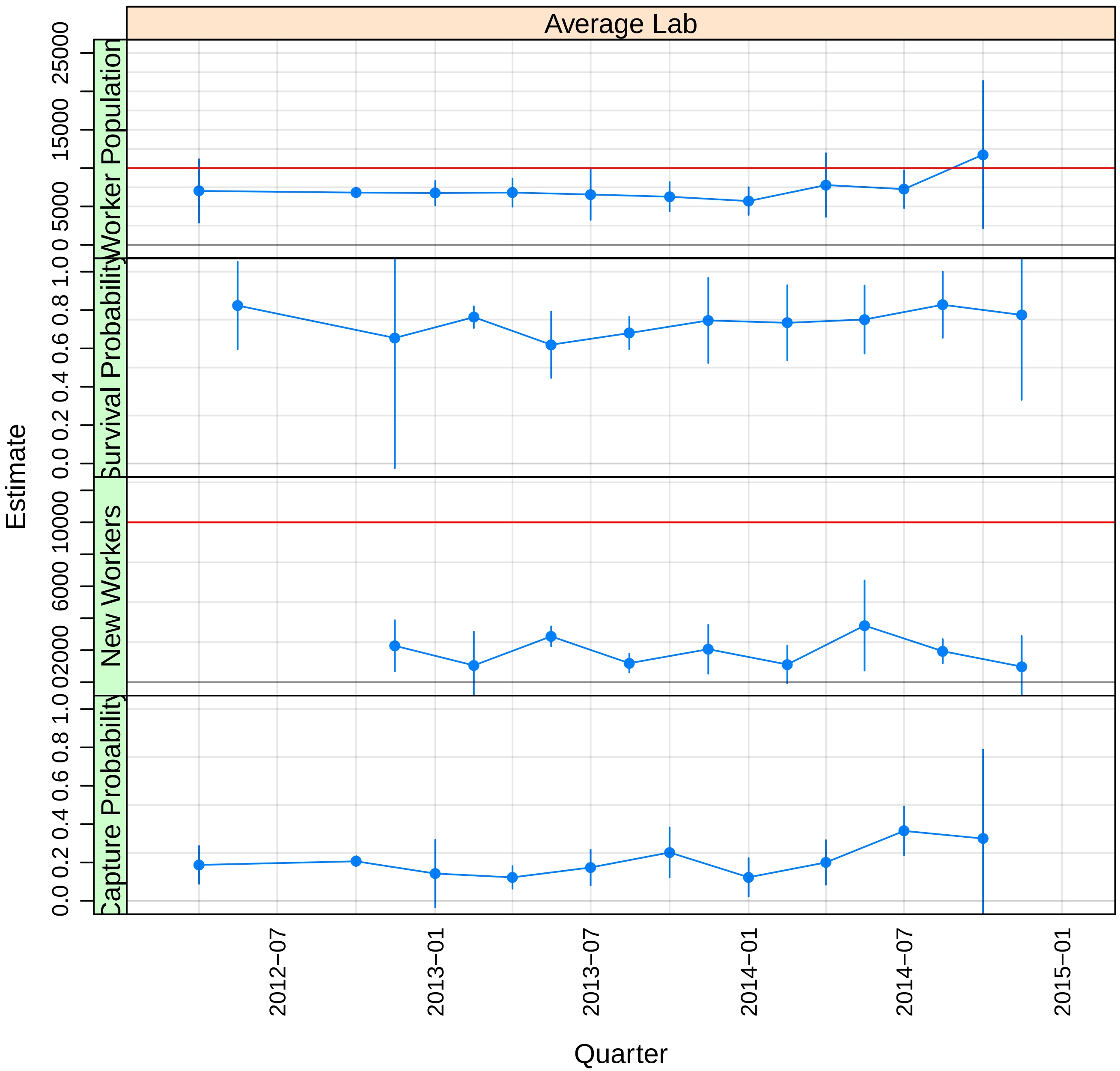
```
## period 2     5018.3   347.5
## period 3     5401.5   565.2
## period 4     6868.4   386.0
## period 5     7368.2   365.7
## period 6     8858.4   430.3
## period 7     6195.4   373.5
## period 8     9818.3   315.8
## period 9     6526.5   200.0
## period 10   10539.3   405.4
## period 11    9755.7   269.8
## period 12    7305.6   146.8
## period 13        --      --
##
## Number of new arrivals:
##                  estimate  stderr
## period 1 -> 2          --      --
## period 2 -> 3      2804.9   561.4
## period 3 -> 4      1466.9   626.4
## period 4 -> 5      2162.8   385.7
## period 5 -> 6      3085.5   403.5
## period 6 -> 7       666.3   406.6
## period 7 -> 8      3622.9   433.9
## period 8 -> 9        29.6   222.1
## period 9 -> 10     5087.7   354.2
## period 10 -> 11    2002.9   298.5
## period 11 -> 12    1955.4   152.7
## period 12 -> 13        --      --
##
## Total number of units who ever inhabited the survey area:
##              estimate  stderr
## all periods   29416.9   321.4
##
## Total number of captured units: 17633
```

```
op.df <- rbindlist(lapply(op, FUN = openp.df))

combineLimits(useOuterStrips(segplot(period ~ lower + upper | lab + type, centers = estimate, data = op.df, horizontal = FALSE, xlab = "Quarter",
  ylab = "Estimate", scales = list(y = list(relation = "free"), x = list(rot = 90), alternating = FALSE), ylim = rep(list(c(0, 1), c(0, 12000),
    c(0, 1), c(0, 20000)), each = 7), xlim = time.range, type = "b")) + layer(panel.abline(h = c(seq(0, 1, 0.2), seq(0, 25000, 2500)), alpha = 0.1)) +
  layer(panel.abline(v = qs, alpha = 0.1)) + layer(panel.abline(h = 10000, col = "red")))
```


Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ungemach
Worker Population
Survival Probability
New Workers
Capture Probability
Estimate
Quarter
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
2012−07
2013−01
2013−07
2014−01
2014−07
2015−01

```r
# Meta analysis to estimate for the average lab

# Just do the meta analysis for one estimate
workerpop <- op.df[type == "Worker Population"]
(ma1 <- rma(y = estimate, sei = stderr, data = workerpop[period == "2013-01-01"]))

##
## Random-Effects Model (k = 2; tau^2 estimator: REML)
##
## tau^2 (estimated amount of total heterogeneity): 903373.9582 (SE = 1947869.2096)
## tau (square root of estimated tau^2 value):      950.4599
## I^2 (total heterogeneity / total variability):   65.59%
## H^2 (total variability / sampling variability):  2.91
##
## Test for Heterogeneity:
## Q(df = 1) = 2.9059, p-val = 0.0883
##
## Model Results:
##
##  estimate        se      zval      pval     ci.lb     ci.ub
## 6743.3009  804.1500    8.3856    <.0001 5167.1958 8319.4060       ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Now do the meta analysis for all estimates
do.rma <- function(data) {
  if (sum(!is.na(data$estimate)) > 1) {
    # Only do rma() on data with at least 2 non-NA observations
    ma1 <- rma(yi = estimate, sei = stderr, data = data)
    data.frame(type = data$type[1], period = data$period[1], estimate = ma1$b, se = ma1$se, I2 = ma1$I2)
  } else NULL
}
# Use by() to run the random-effects meta analysis for each statistic for each period
estimates <- by(data = op.df, INDICES = list(op.df$type, op.df$period), do.rma)

## Warning in rma(yi = estimate, sei = stderr, data = data):  Studies with NAs omitted from model fitting.

estimates <- rbindlist(estimates)

# Median heterogeneity estimate
median(estimates[type == "Worker Population"]$I2)

## [1] 95.50962

# Add 95% CIs
```

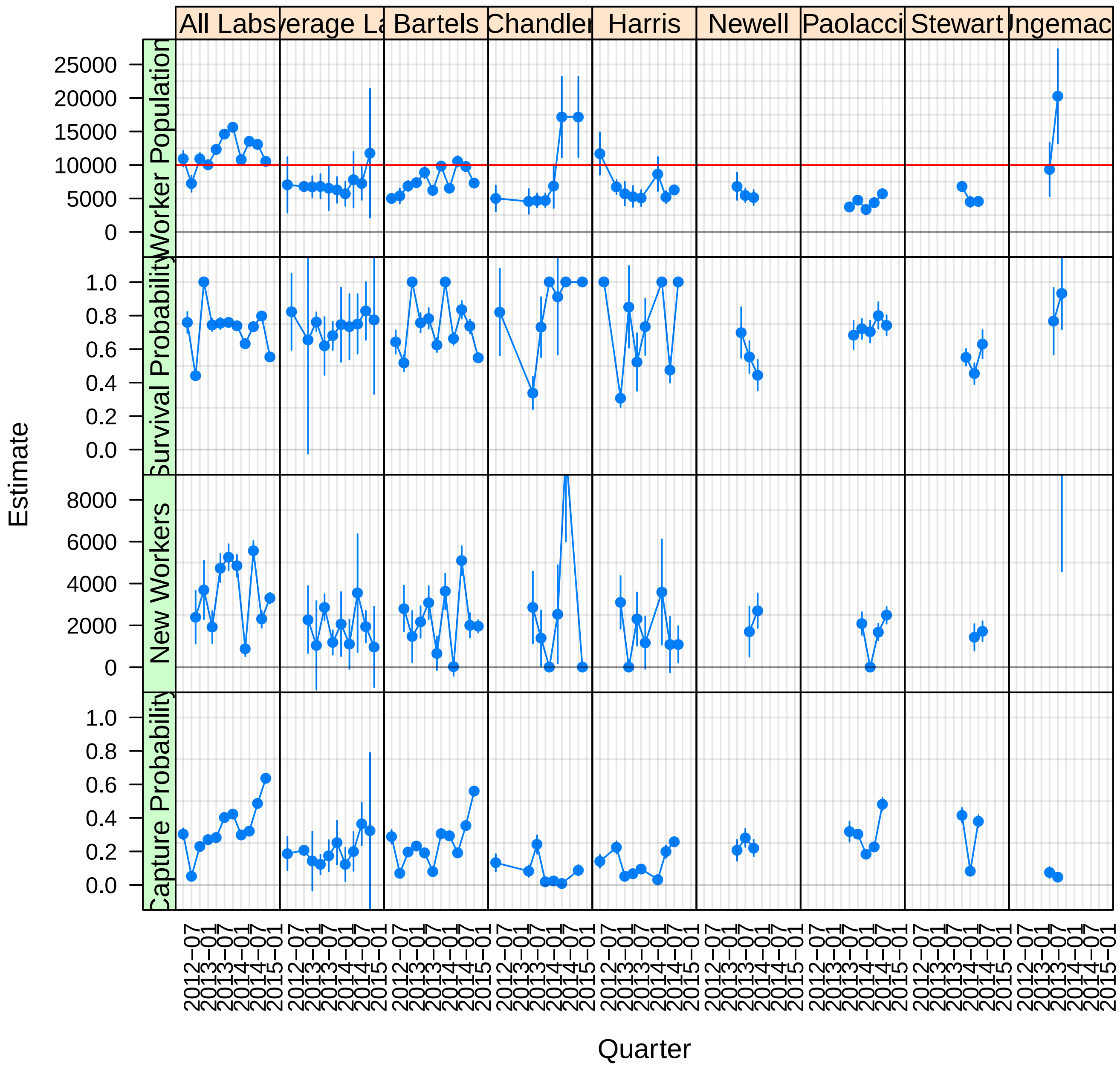
```r
estimates$lower <- with(estimates, estimate - qnorm(0.975) * se)
estimates$upper <- with(estimates, estimate + qnorm(0.975) * se)
estimates$type <- as.character(estimates$type)

useOuterStrips(segplot(period ~ lower + upper | "Average Lab" + type, centers = estimate, data = estimates, horizontal = FALSE, xlab = "Quarter",
  ylab = "Estimate", scales = list(y = list(relation = "free"), x = list(rot = 90), alternating = FALSE), type = "b", ylim = list(c(0, 1),
    c(0, 12000), c(0, 1), c(0, 25000)), xlim = time.range)) + layer(panel.abline(h = c(seq(0, 1, 0.25), seq(0, 25000, 2500)), alpha = 0.1)) +
  layer(panel.abline(v = qs, alpha = 0.1)) + layer(panel.abline(h = 10000, col = "red"))
```



Average Lab
Worker Population
Survival Probability
New Workers
Capture Probability
Estimate
0 5000 15000 25000
0.0 0.2 0.4 0.6 0.8 1.0
0 2000 6000 10000
0.0 0.2 0.4 0.6 0.8 1.0
2012−07
2013−01
2013−07
2014−01
2014−07
2015−01
Quarter

```r
estimates$lab <- "Average Lab"
```

```r
# Plot all together for Figure 2
all.combined <- rbind(openp.df(op.all), op.df, estimates, fill = TRUE)

(all.data.op.plot <- combineLimits(useOuterStrips(segplot(period ~ lower + upper | lab + type, centers = estimate, data = all.combined, horizontal = FALSE,
  xlab = "Quarter", ylab = "Estimate", scales = list(y = list(relation = "free", rot = 0), x = list(rot = 90), alternating = FALSE), ylim = rep(list(c(0,
    1), c(0, 8000), c(0, 1), c(0, 25000)), each = 9), xlim = time.range, type = "b")) + layer(panel.abline(h = c(seq(0, 1, 0.25), seq(0,
  25000, 2500)), alpha = 0.1)) + layer(panel.abline(v = qs, alpha = 0.1)) + layer(panel.abline(h = 10000, col = "red"))))
```



All Labs
verage La
Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ingemac
Worker Population
Survival Probability
New Workers
Capture Probability
Estimate
Quarter

```
pdf("open_population_plot.pdf", width = 12, height = 8)
all.data.op.plot
dev.off()
```

```
## pdf
##   2
```

```
# Means over time Includes the mean over time of the population estimate for the average lab headlined in the title of the paper
(est <- estimates[, .(mean.over.time = mean(estimate)), by = type])
```

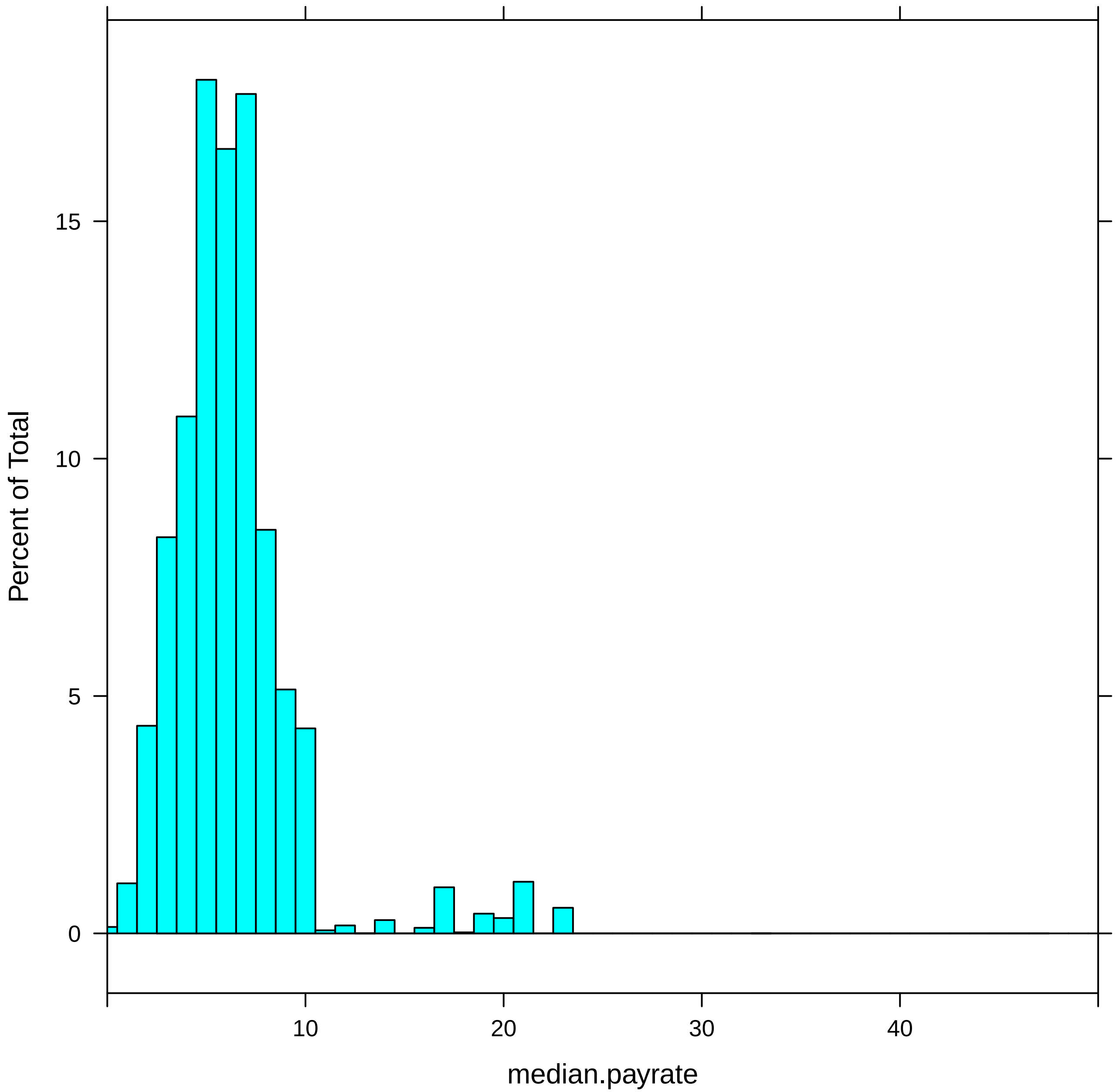
```
##                    type mean.over.time
## 1:  Capture Probability      0.2096507
## 2:    Worker Population   7272.7290270
## 3: Survival Probability      0.7372017
## 4:          New Workers   1888.6442290
```

```
mean.survival.prob <- est[type == "Survival Probability"]$mean.over.time
# Half life in months
log(0.5)/log(mean.survival.prob)/4 * 12
```

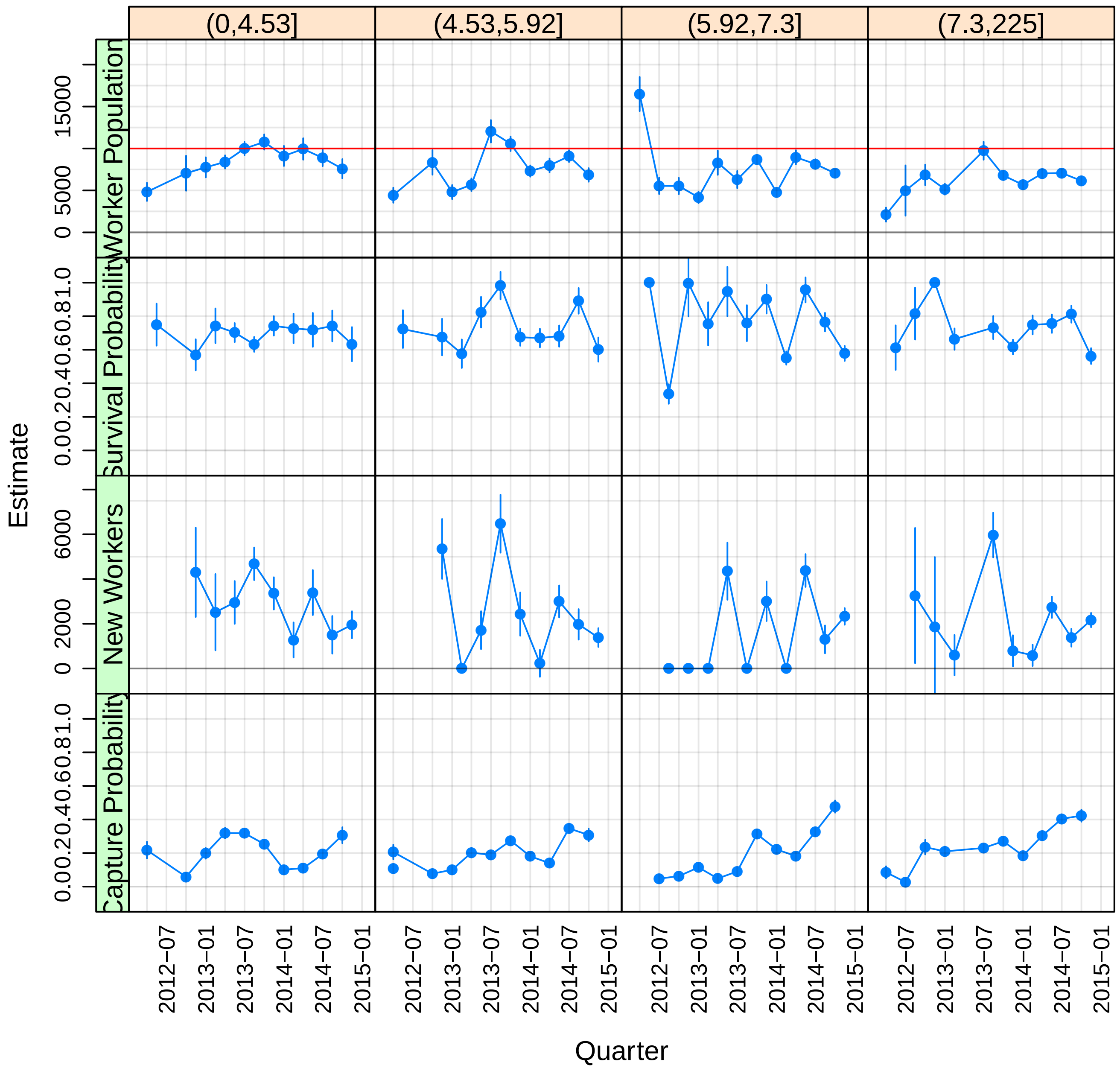
```
## [1] 6.820217
```

4 Section 3.1: Pay

```
histogram(~median.payrate, data = HITs, breaks = 0:1000 - 0.5, xlim = c(0, 50))
```


Percent of Total
0
5
10
15
10
20
30
40
median.payrate

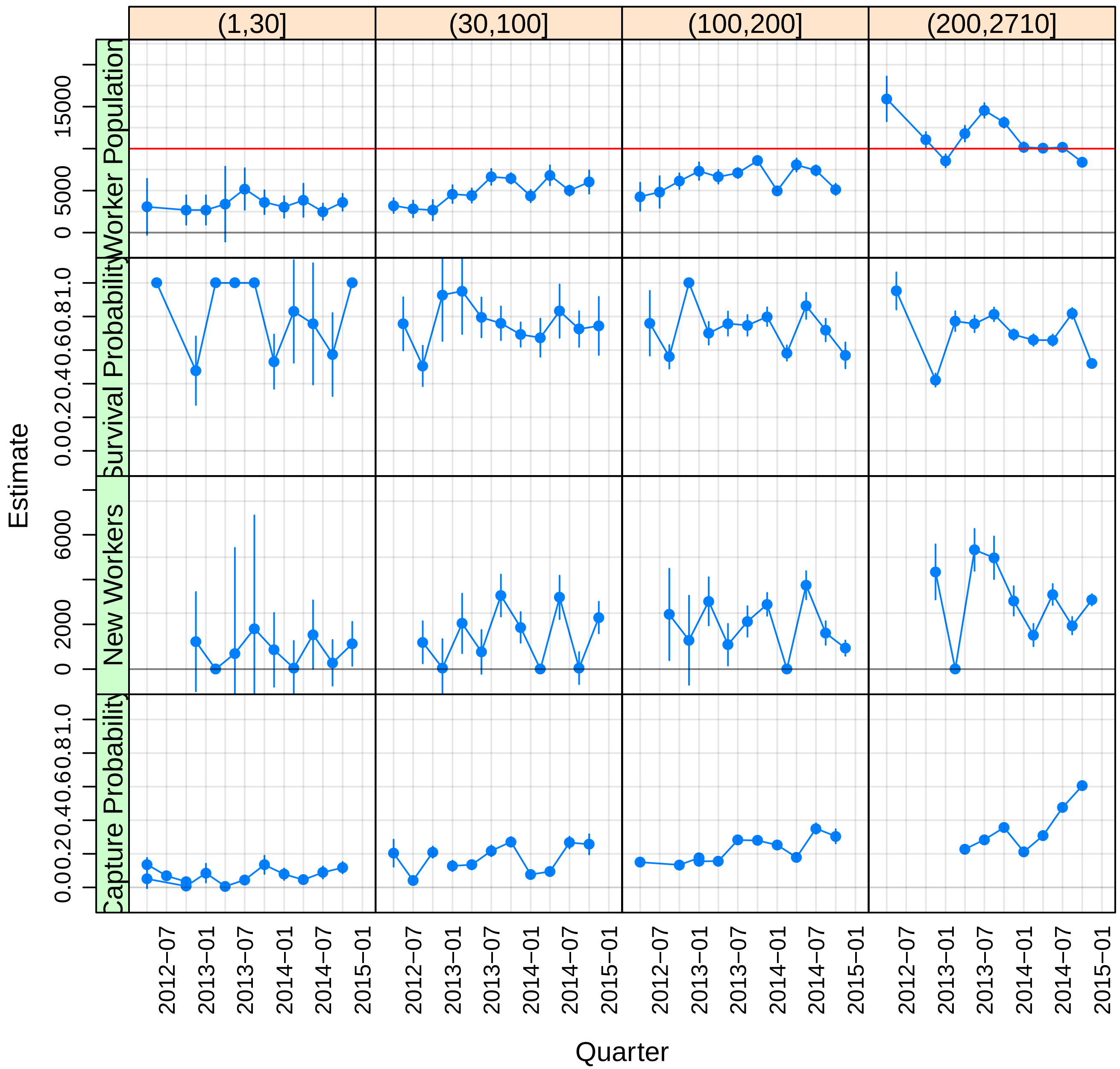
```
HITs <- HITs[, `:=`(pay.rate.quantile, cut(median.payrate, quantile(HITs$median.payrate)))]
op <- by(HITs, INDICES = list(HITs$pay.rate.quantile), FUN = cap.recap.openp, lab = "get.from.HITs.data.table")
op.df <- rbindlist(lapply(op, FUN = openp.df))

op.df$lab <- rep(levels(HITs$pay.rate.quantile), each = 47)

(pay.openp <- combineLimits(useOuterStrips(segplot(period ~ lower + upper | lab + type, centers = estimate, data = op.df, horizontal = FALSE,
  xlab = "Quarter", ylab = "Estimate", scales = list(y = list(relation = "free"), x = list(rot = 90), alternating = FALSE), ylim = rep(list(c(0,
    1), c(0, 7500), c(0, 1), c(0, 20000)), each = 4), xlim = time.range, type = "b")) + layer(panel.abline(h = c(seq(0, 1, 0.2), seq(0, 25000,
  2500)), alpha = 0.1)) + layer(panel.abline(v = qs, alpha = 0.1)) + layer(panel.abline(h = 10000, col = "red"))))
```


(0,4.53]
(4.53,5.92]
(5.92,7.3]
(7.3,225]
Worker Population
Survival Probability
New Workers
Capture Probability
Estimate
Quarter
0
5000
15000
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
6000
2012−07
2013−01
2013−07
2014−01
2014−07
2015−01

```
# Figure 3
pdf("open_population_by_pay.pdf", width = 8, height = 8)
pay.openp
dev.off()

## pdf
##   2
```

5 Section 3.2: Batch Size

```
batch.size <- HITs[, .(batch.size = .N), by = filename]
batch.size.quantiles <- quantile(batch.size$batch.size)

HITs <- merge(HITs, batch.size, by = "filename")

HITs <- HITs[, `:=`(batch.size.quantile, cut(batch.size, batch.size.quantiles))]
op <- by(HITs, INDICES = list(HITs$batch.size.quantile), FUN = cap.recap.openp, lab = "get.from.HITs.data.table")
op.df <- rbindlist(lapply(op, FUN = openp.df))

op.df$lab <- rep(levels(HITs$batch.size.quantile), each = 50)
op.df$lab <- factor(op.df$lab, levels = levels(HITs$batch.size.quantile), labels = c("(1,30]", "(30,100]", "(100,200]", "(200,2710]"))
# Figure 4
(batch.size.openp <- combineLimits(useOuterStrips(segplot(period ~ lower + upper | lab + type, centers = estimate, data = op.df, horizontal = FALSE,
  xlab = "Quarter", ylab = "Estimate", scales = list(y = list(relation = "free"), x = list(rot = 90), alternating = FALSE), ylim = rep(list(c(0,
    1), c(0, 7500), c(0, 1), c(0, 20000)), each = 4), xlim = time.range, type = "b")) + layer(panel.abline(h = c(seq(0, 1, 0.2), seq(0, 25000,
  2500)), alpha = 0.1)) + layer(panel.abline(v = qs, alpha = 0.1)) + layer(panel.abline(h = 10000, col = "red"))))
```



(1,30]
(30,100]
(100,200]
(200,2710]
Worker Population
Survival Probability
New Workers
Capture Probability
Estimate
Quarter
0
5000
15000
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
6000
2012−07
2013−01
2013−07
2014−01
2014−07
2015−01

```
pdf("open_population_by_batch_size.pdf", width = 8, height = 8)
batch.size.openp
dev.off()
```

```
## pdf
##   2
```

```
op.df[type == "Worker Population", mean(estimate, na.rm = TRUE), by = lab]
```

```
##            lab        V1
## 1:      (1,30]  3368.251
## 2:    (30,100]  4820.588
## 3:   (100,200]  6405.137
## 4: (200,2710] 11376.712
```

```
# 95% CIs for averages over time
op.df.av <- op.df[type == "Worker Population", .(estimate = mean(estimate, na.rm = TRUE), stderr = sqrt(sum(stderr^2, na.rm = TRUE))/sum(!is.na(stderr))),
  by = lab]
op.df.av <- op.df.av[, `:=`(lower.CI, estimate - qnorm(0.975) * stderr)]
op.df.av <- op.df.av[, `:=`(upper.CI, estimate + qnorm(0.975) * stderr)]
op.df.av
```

```
##            lab  estimate   stderr  lower.CI  upper.CI
## 1:      (1,30]  3368.251 371.7662  2639.603  4096.899
## 2:    (30,100]  4820.588 152.6693  4521.362  5119.815
## 3:   (100,200]  6405.137 156.1569  6099.075  6711.199
## 4: (200,2710] 11376.712 173.3011 11037.048 11716.376
```

6 Section 3.3: Robustness of the Open Population Estimate

```
# Keeping only people caught fewer than 10 times
sum(op.all$descriptive$base.freq[, "ui"][10:13])/op.all$descriptive$n  # Proportion of workers caught more than 10 times
```

```
## [1] 0.3363428
```

```
keep <- apply(histpos.t(13), 1, sum) < 10
# Run open-population analysis only with workers caught fewer than 10 times
op.all.fewer.than.10 <- cap.recap.openp(HITs, lab = "All", keep = keep)

op.all$openp$N[, "estimate"]
```

```
##   period 1  period 2  period 3  period 4  period 5  period 6  period 7  period 8  period 9 period 10 period 11 period 12
##         NA 10939.054  7219.155 10912.625 10040.805 12305.603 14593.036 15646.704 10804.628 13508.548 13068.137 10546.555
## period 13
```

```
##        NA
mean(op.all$openp$N[, "estimate"], na.rm = TRUE)

## [1] 11780.44

mean(op.all$openp$birth[, "estimate"], na.rm = TRUE)

## [1] 3490.102

op.all.fewer.than.10$openp$N[, "estimate"]

##  period 1  period 2  period 3  period 4  period 5  period 6  period 7  period 8  period 9 period 10 period 11 period 12
##        NA 11989.901  8475.239 12157.962 10813.343 12981.520 15012.733 15995.991 11123.208 13883.437 13286.252 10667.496
## period 13
##        NA

mean(op.all.fewer.than.10$openp$N[, "estimate"], na.rm = TRUE)

## [1] 12398.83
```

```
# US workers with a HIT acceptance rate requirement of greater than 80%
HITs <- HITs.original
# As before, but also only UNITED STATES and high HIT requirements
HITs <- HITs[conditional == "Open" & multiple.responses == "No" & location.requirement == "UNITED STATES" & HIT.requirement > 50]
# Fraction remaining compared to original analysis
nrow(HITs)/nrow(HITs.original)

## [1] 0.7331819

(op.all <- cap.recap.openp(HITs, lab = "All Labs"))

## $periods
##  [1] "2012-01-01" "2012-04-01" "2012-07-01" "2012-10-01" "2013-01-01" "2013-04-01" "2013-07-01" "2013-10-01"
##  [9] "2014-01-01" "2014-04-01" "2014-07-01" "2014-10-01" "2015-01-01"
##
## $descriptive
##
## Number of captured units: 28672
##
## Frequency statistics:
##          fi     ui     vi     ni
## i = 1   19592   1828   1134   1828
## i = 2    4807   2888   2437   3326
## i = 3    1894    277    128    378
## i = 4     918   2014   1375   2518
## i = 5     588   1871   1389   2718
## i = 6     362   2424   1757   3493
```

```
## i = 7     225   3030   2366   4744
## i = 8     116   3346   3171   5666
## i = 9      79   1183   1495   3043
## i = 10     55   2283   1705   3972
## i = 11     23   3164   3648   6173
## i = 12      7   3097   4385   6310
## i = 13      6   1267   3682   3682
## fi: number of units captured i times
## ui: number of units captured for the first time on occasion i
## vi: number of units captured for the last time on occasion i
## ni: number of units captured on occasion i
##
##
## $lab
## [1] "All Labs"
##
## $openp
##
## Model fit:
##               deviance      df        AIC
## fitted model  10371.32    8156   13773.35
##
## Test for trap effect:
##                                   deviance      df        AIC
## model with homogenous trap effect  8605.428    8155   12009.46
## model with trap effect             8538.029    8146   11960.06
##
## Capture probabilities:
##            estimate  stderr
## period 1         --      --
## period 2     0.3138  0.0180
## period 3     0.0540  0.0054
## period 4     0.2348  0.0105
## period 5     0.2761  0.0100
## period 6     0.2830  0.0090
## period 7     0.3858  0.0093
## period 8     0.4174  0.0090
## period 9     0.3146  0.0079
## period 10    0.3343  0.0080
## period 11    0.5110  0.0094
## period 12    0.6667  0.0116
## period 13        --      --
##
## Survival probabilities:
```

```
##                  estimate  stderr
## period 1 -> 2      0.7635  0.0333
## period 2 -> 3      0.4364  0.0145
## period 3 -> 4      1.0000  0.0000
## period 4 -> 5      0.7374  0.0189
## period 5 -> 6      0.7648  0.0184
## period 6 -> 7      0.7165  0.0145
## period 7 -> 8      0.7438  0.0136
## period 8 -> 9      0.6639  0.0133
## period 9 -> 10     0.7123  0.0146
## period 10 -> 11    0.8027  0.0138
## period 11 -> 12    0.5323  0.0102
## period 12 -> 13        --      --
##
## Abundances:
##            estimate  stderr
## period 1         --      --
## period 2    10598.9   589.7
## period 3     6997.0   605.1
## period 4    10724.3   439.9
## period 5     9845.2   317.0
## period 6    12342.4   350.5
## period 7    12297.6   259.6
## period 8    13574.9   256.7
## period 9     9671.2   194.5
## period 10   11883.1   240.5
## period 11   12081.2   194.5
## period 12    9464.4   149.2
## period 13        --      --
##
## Number of new arrivals:
##                  estimate  stderr
## period 1 -> 2          --      --
## period 2 -> 3      2371.3   625.7
## period 3 -> 4      3727.3   694.7
## period 4 -> 5      1937.0   383.1
## period 5 -> 6      4813.0   349.8
## period 6 -> 7      3454.5   289.2
## period 7 -> 8      4428.1   243.9
## period 8 -> 9       659.1   173.8
## period 9 -> 10     4994.8   222.3
## period 10 -> 11    2542.1   198.1
## period 11 -> 12    3033.1   121.5
## period 12 -> 13        --      --
```

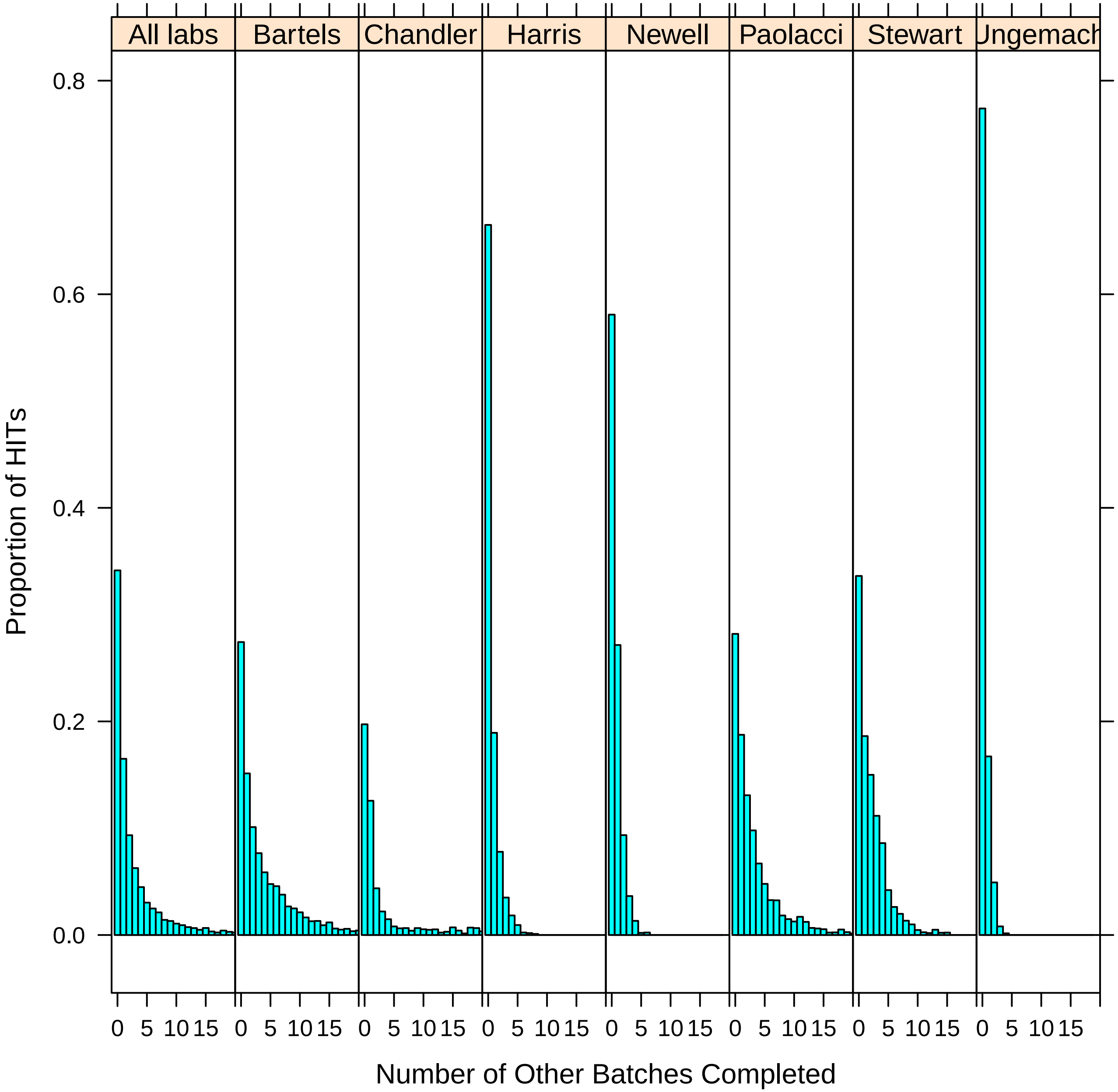
```
##
## Total number of units who ever inhabited the survey area:
##               estimate  stderr
## all periods   43786.3   406.9
##
## Total number of captured units: 28672
```

```
mean(op.all$openp$N[, "estimate"], na.rm = TRUE)
```

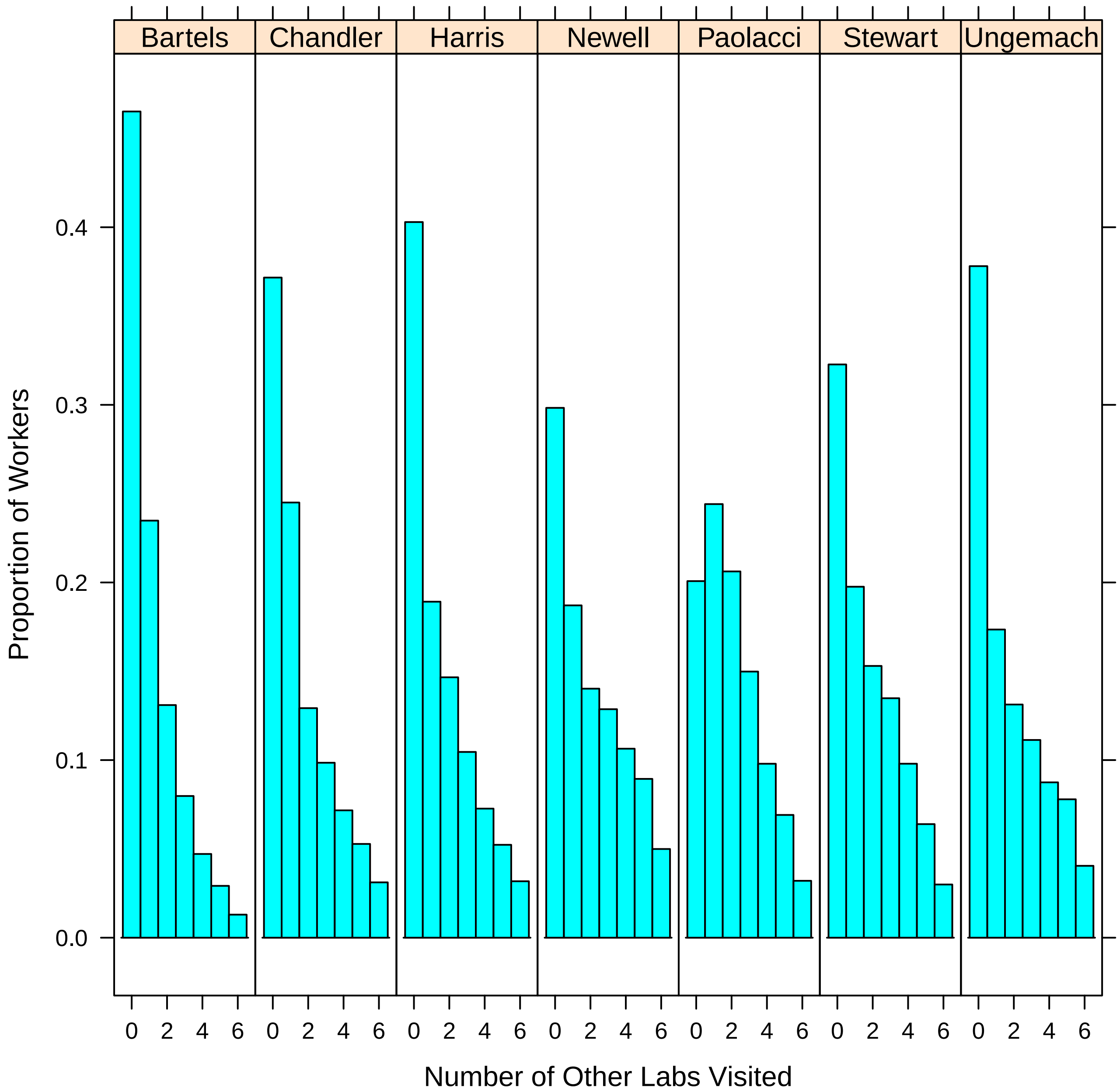
```
## [1] 10861.84
```

7 Repeated Participation

```
HITs <- HITs.original

# The distribution of the number of other batches completed within a laboratory

# Add a column to HITs for the number of batches completed by each worker
HITs <- HITs[, `:=`(N.batches, .N), by = .(WorkerId)]
# ... and for within each lab
HITs <- HITs[, `:=`(N.batches.within.lab, .N), by = .(WorkerId, lab)]

HITs.all.labs <- HITs
HITs.all.labs$lab <- "All labs"
HITs.all.labs <- rbind(HITs, HITs.all.labs)

# Figure 5
(no.batches.plot <- histogram(~(N.batches.within.lab - 1) | lab, breaks = (-1):1000 + 0.5, xlim = c(-1, 20), data = HITs.all.labs, scales = list(alternating
  as.table = TRUE, layout = c(8, 1), xlab = "Number of Other Batches Completed", ylab = "Proportion of HITs", type = "density"))
```

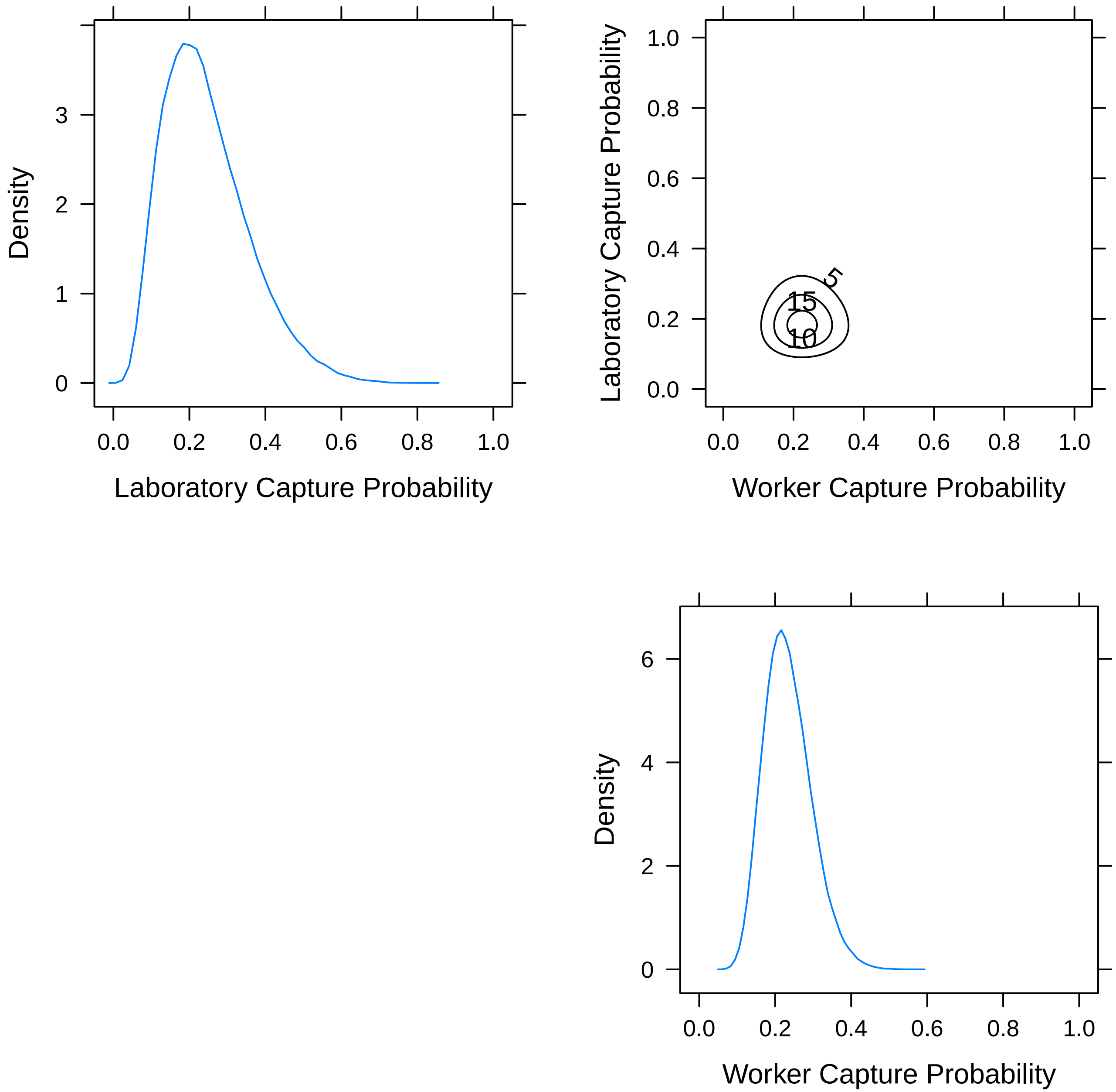

All labs
Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ungemach
Proportion of HITs
0.0
0.2
0.4
0.6
0.8
0 5 10 15
Number of Other Batches Completed

```
pdf("no_batches_plot.pdf", width = 12, height = 4)
no.batches.plot
dev.off()

## pdf
##   2

round(prop.table(xtabs(~N.batches.within.lab, data = HITs.all.labs[lab == "Bartels"])), digits = 2)

## N.batches.within.lab
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24
## 0.27 0.15 0.10 0.08 0.06 0.05 0.05 0.04 0.03 0.02 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.00 0.00 0.00 0.00 0.00
##   25   26   27   28   29   30   31   32   33   34   36   37   40   43   48   53   57   61   65   73   84   85
## 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00

round(prop.table(xtabs(~N.batches.within.lab, data = HITs.all.labs[lab == "All labs"])), digits = 2)

## N.batches.within.lab
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24
## 0.34 0.17 0.09 0.06 0.04 0.03 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45   47   48   49
## 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##   50   51   52   53   54   55   56   57   59   60   61   62   63   65   67   68   69   70   72   73   74   77   80   82
## 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##   84   85   89   94   95   96   97   99  100  101  104  108  109  113  116  117  121  132  139  149  187  228  231  232
## 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
##  233  234  240  251  254  312  333  430  450  636
## 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01

# The distribution of the number of other laboratories visited

# No number of other labs participated in
workers.by.lab <- HITs[, WorkerId, by = .(WorkerId, lab)]
workers.by.lab <- workers.by.lab[, `:=`(N, .N), by = WorkerId]

# Figure 6
(no.labs.plot <- histogram(~(N - 1) | lab, data = workers.by.lab, type = "density", breaks = (-1):6 + 0.5, layout = c(7, 1), xlab = "Number of Other Labs Vi
  ylab = "Proportion of Workers", scales = list(alternating = FALSE)))
```


Bartels
Chandler
Harris
Newell
Paolacci
Stewart
Ungemach
Proportion of Workers
0.0
0.1
0.2
0.3
0.4
0 2 4 6
Number of Other Labs Visited

```
pdf("no_labs_plot.pdf", width = 12, height = 4)
no.labs.plot
dev.off()

## pdf
##   2

# The joint distribution of worker and laboratory capture probabilities, together with marginal distributions
HITs <- HITs.original

lab.by.worker <- xtabs(~WorkerId + lab, data = HITs)
lab.by.worker[lab.by.worker > 1] <- 1
freqs <- melt(lab.by.worker)
# Select a random sample of 100 workers for modelling, which means results will vary from the sample in the paper
selected.workers <- sample(unique(freqs$WorkerId), 100)
selected.freqs <- droplevels(subset(freqs, WorkerId %in% selected.workers))

mm1 <- glmer(value ~ (1 | lab) + (1 | WorkerId), data = selected.freqs, family = binomial)
summary(mm1)

## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: value ~ (1 | lab) + (1 | WorkerId)
##    Data: selected.freqs
##
##      AIC      BIC   logLik deviance df.resid
##    749.5    763.1   -371.7    743.5      697
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1511 -0.5171 -0.4719 -0.3933  2.5428
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  WorkerId (Intercept) 0.1247   0.3532
##  lab      (Intercept) 0.3998   0.6323
## Number of obs: 700, groups:  WorkerId, 100; lab, 7
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2188     0.2612  -4.667 3.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

x <- mvrnorm(1e+05, rep(fixef(mm1), 2), diag(VarCorr(mm1)))
```

```r
logit <- function(x) {
  1/(1 + exp(-x))
}
x <- logit(x)
z <- kde2d(x = x[, 1], y = x[, 2], h = c(0.2, 0.2), n = 100)

joint.plot <- contourplot(z$z, row.values = z$x, column.values = z$x, xlim = c(-0.05, 1.05), ylim = c(-0.05, 1.05), xlab = "Worker Capture Probability",
  ylab = "Laboratory Capture Probability")
worker.plot <- densityplot(~x[, 1], plot.points = FALSE, xlab = "Worker Capture Probability", xlim = c(-0.05, 1.05))
lab.plot <- densityplot(~x[, 2], plot.points = FALSE, xlab = "Laboratory Capture Probability", xlim = c(-0.05, 1.05))

# Figure 7
plot(worker.plot, split = c(2, 2, 2, 2))
plot(joint.plot, split = c(2, 1, 2, 2), newpage = FALSE)
plot(lab.plot, split = c(1, 1, 2, 2), newpage = FALSE)
```


Density
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
Laboratory Capture Probability
Laboratory Capture Probability
0.0
0.2
0.4
0.6
0.8
1.0
5
15
10
0.0
0.2
0.4
0.6
0.8
1.0
Worker Capture Probability
Density
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
Worker Capture Probability

```r
pdf("worker_capture_prob_density.pdf", width = 4, height = 4)
worker.plot
dev.off()
```

```
## pdf
##   2
```

```r
pdf("lab_capture_prob_density.pdf", width = 4, height = 4)
lab.plot
dev.off()
```

```
## pdf
##   2
```

```r
pdf("joint_capture_prob_density.pdf", width = 4, height = 4)
joint.plot
dev.off()
```

```
## pdf
##   2
```

```r
logit(fixef(mm1)[1])
```

```
## (Intercept)
##   0.2281418
```

```r
quantile(x[, 1], c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1290969 0.3710997
```

```r
quantile(x[, 2], c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.07877207 0.50594362
```