

Litter size and fertility: Learning from the data

Data

Parsing (PDF → table)

See here for PDF to table conversion and parsing.

Load and clean the data

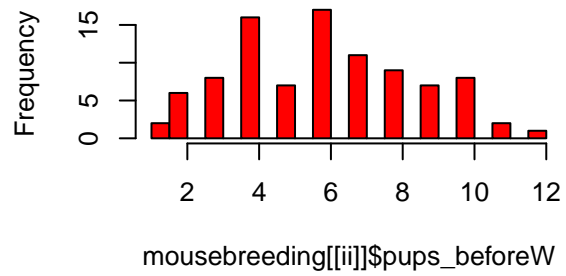
First of all, we load the litter size data (the number of pups born etc.) per strain from the already generated txt files, and then do a small quality check: Look at the distributions of the litter size (it should look like a Poisson, or at least be one-modal).

```
names(mousebreeding)
```

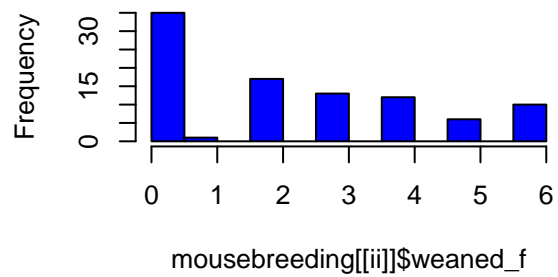
```
## [1] "B6cBrd"      "B6D2F1"      "B6J_Cr1F"    "B6J_Fue"
## [5] "Balbc"       "Card9_KO"    "CD1_1999_2010" "CD1_2010_2020"
## [9] "DBA2_J_Fue"  "FcRn"        "NMRI"
```

```
for (ii in 1:length(mousebreeding)){
  par(mfrow=c(2,2))
  hist(mousebreeding[[ii]]$pups_born, breaks = 20, col = "orange")
  hist(mousebreeding[[ii]]$pups_beforeW, breaks = 20, col = "red")
  hist(mousebreeding[[ii]]$weaned_f, breaks = 20, col = "pink")
  hist(mousebreeding[[ii]]$weaned_f, breaks = 20, col = "blue")
}
```

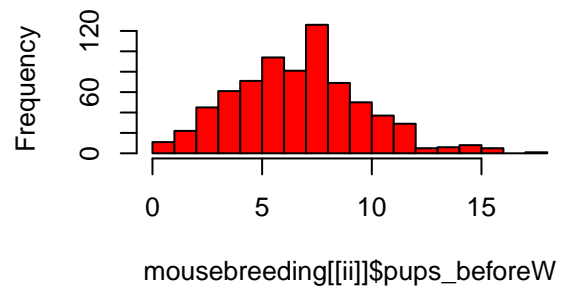
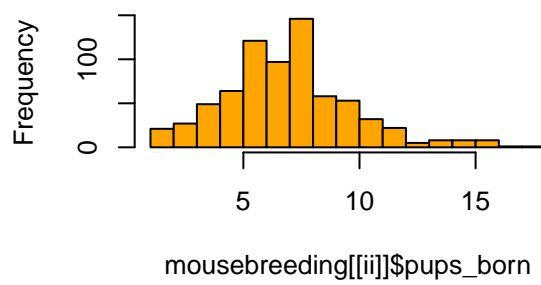
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



Histogram of mousebreeding[[ii]]\$weane Histogram of mousebreeding[[ii]]\$weane



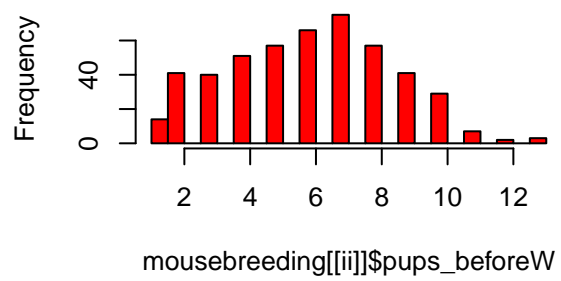
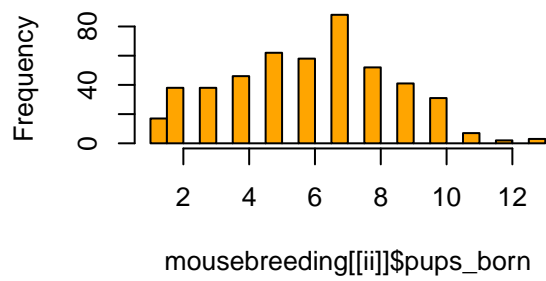
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



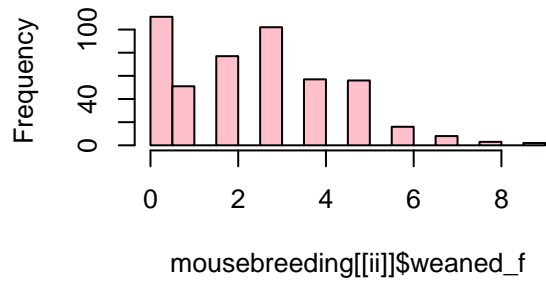
Histogram of mousebreeding[[ii]]\$weane Istogram of mousebreeding[[ii]]\$weane



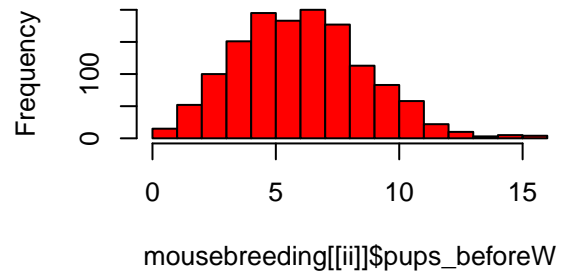
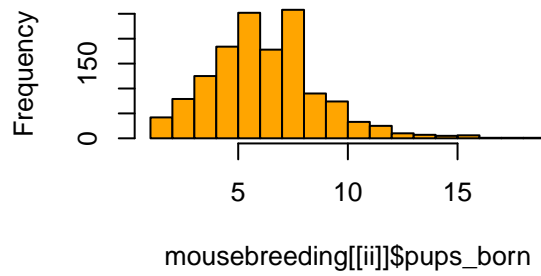
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



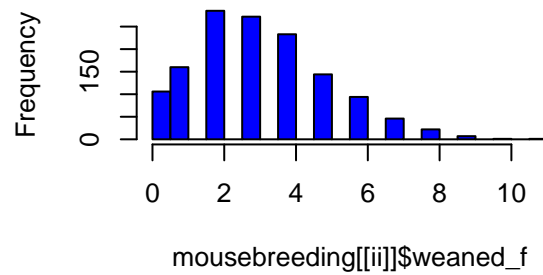
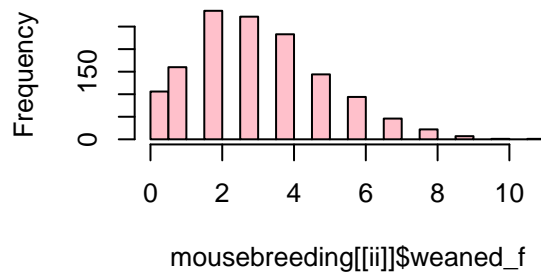
Histogram of mousebreeding[[ii]]\$weane Histogram of mousebreeding[[ii]]\$weane



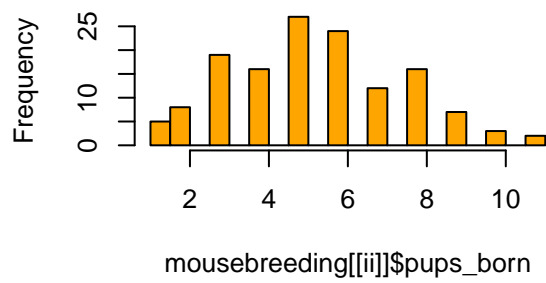
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



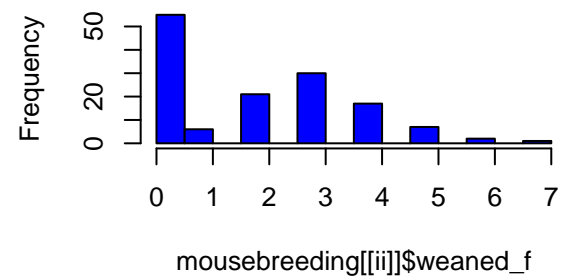
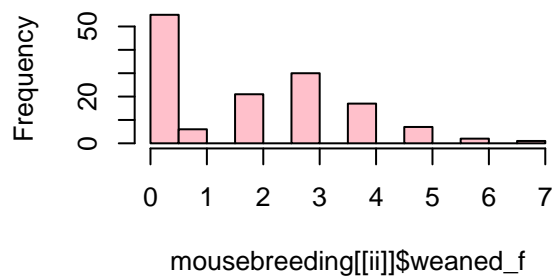
Histogram of mousebreeding[[ii]]\$weane Istogram of mousebreeding[[ii]]\$weane



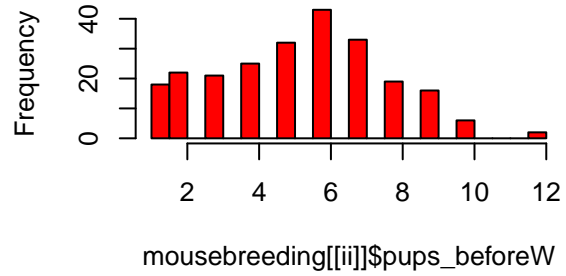
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



Histogram of mousebreeding[[ii]]\$weane Istogram of mousebreeding[[ii]]\$weane



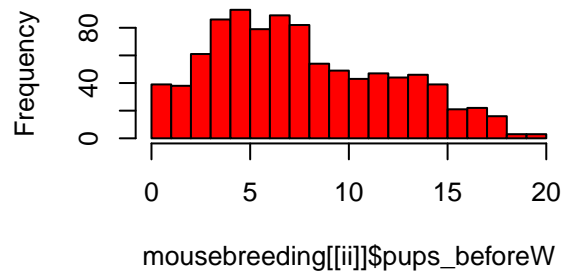
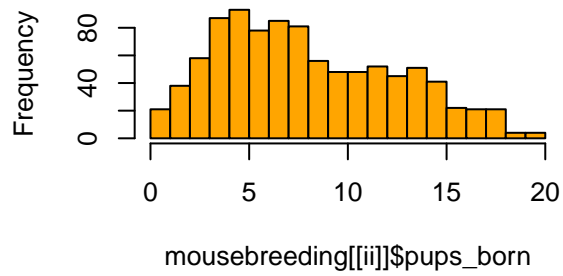
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



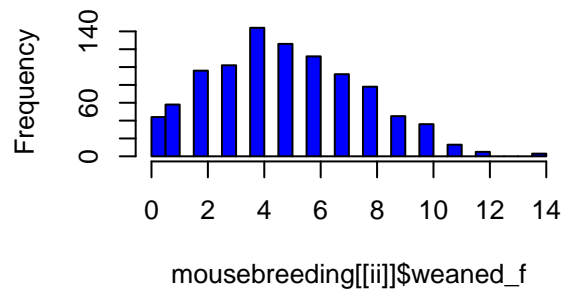
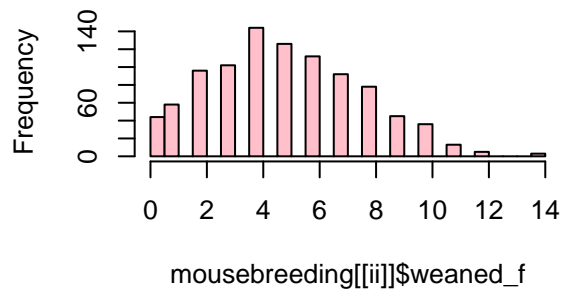
Histogram of mousebreeding[[ii]]\$weane Istogram of mousebreeding[[ii]]\$weane



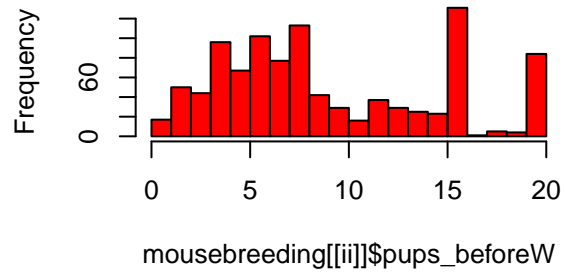
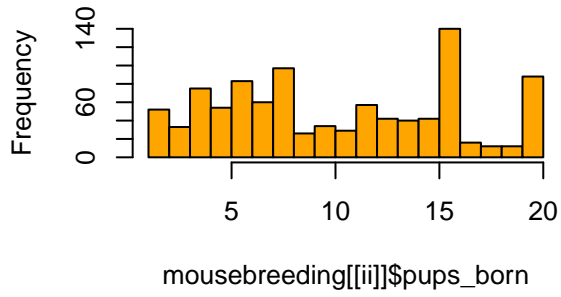
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



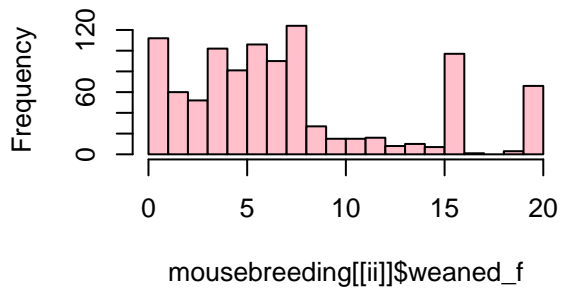
Histogram of mousebreeding[[ii]]\$weane Istogram of mousebreeding[[ii]]\$weane



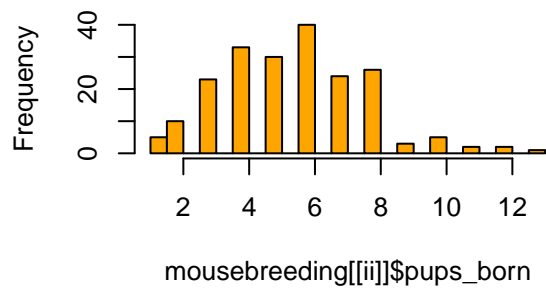
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



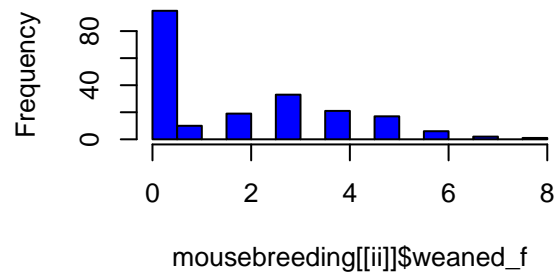
Histogram of mousebreeding[[ii]]\$weane Histogram of mousebreeding[[ii]]\$weane



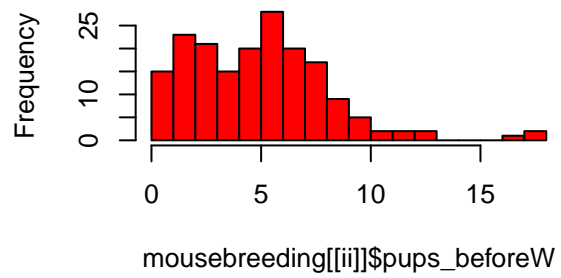
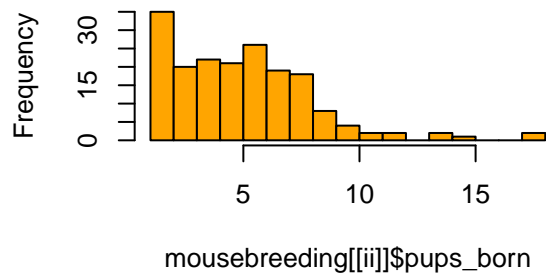
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



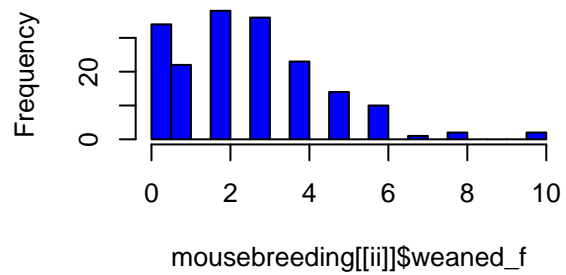
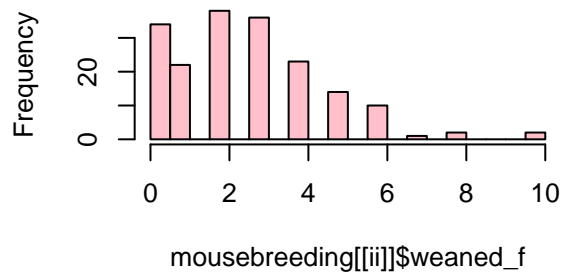
Histogram of mousebreeding[[ii]]\$weane Histogram of mousebreeding[[ii]]\$weane



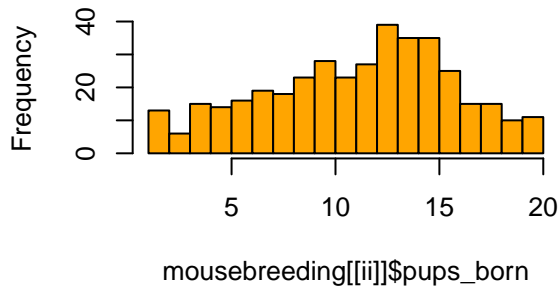
Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



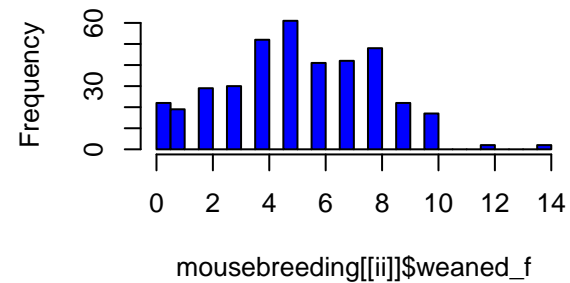
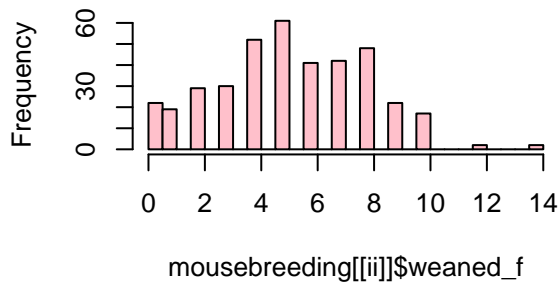
Histogram of mousebreeding[[ii]]\$weane Istogram of mousebreeding[[ii]]\$weane



Histogram of mousebreeding[[ii]]\$pups_ Istogram of mousebreeding[[ii]]\$pups_be



Histogram of mousebreeding[[ii]]\$weane Histogram of mousebreeding[[ii]]\$weane



Quite often, only pups of one particular gender are required for the experiment. Therefore, all pups of the other gender are killed - and this is usually not stated in the data. Therefore, I filter all these cases out. Here is the number of litters per strain before (left) and after cleaning (right).

```
nlitters
```

```
##           before after
## B6cBrd         94    77
## B6D2F1        721   565
## B6J_CrlF       483   444
## B6J_Fue       1371  1219
## Balbc         139   116
## Card9_KO       237   211
## CD1_1999_2010  954   593
## CD1_2010_2020  992   280
## DBA2_J_Fue     204   178
## FcRn           182   145
## NMRI           387   375
```

```
names(table_list_clean)
```

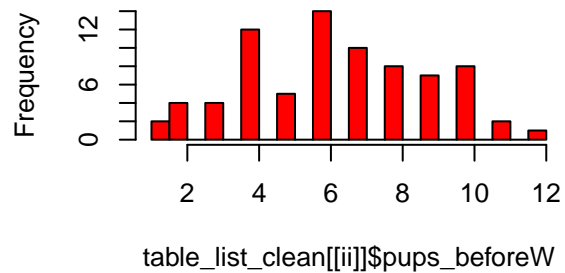
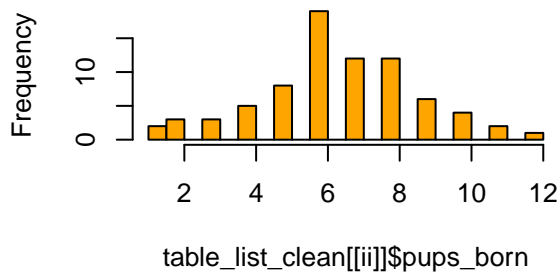
```
## [1] "B6cBrd"      "B6D2F1"      "B6J_CrlF"    "B6J_Fue"
## [5] "Balbc"       "Card9_KO"    "CD1_1999_2010" "CD1_2010_2020"
## [9] "DBA2_J_Fue"  "FcRn"        "NMRI"
```

```

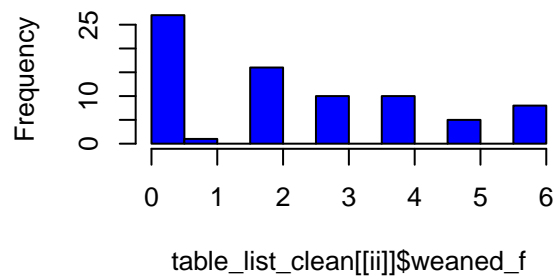
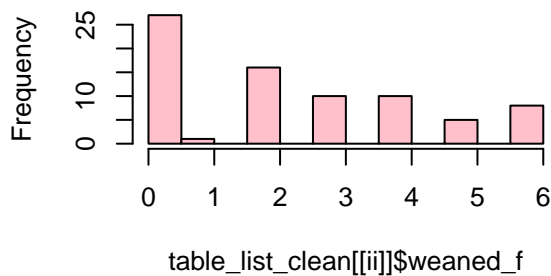
for (ii in 1:length(table_list_clean)){
  par(mfrow=c(2,2))
  hist(table_list_clean[[ii]]$pups_born, breaks = 20, col = "orange")
  hist(table_list_clean[[ii]]$pups_beforeW, breaks = 20, col = "red")
  hist(table_list_clean[[ii]]$weaned_f, breaks = 20, col = "pink")
  hist(table_list_clean[[ii]]$weaned_f, breaks = 20, col = "blue")
}

```

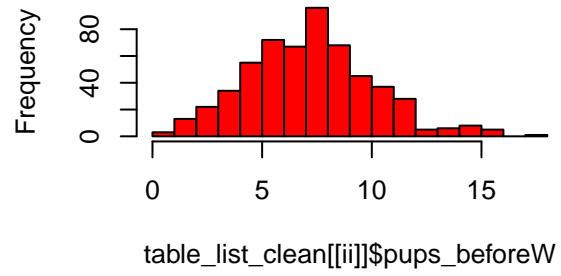
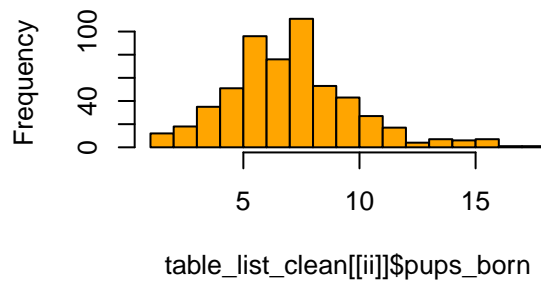
Histogram of table_list_clean[[ii]]\$pups_born Histogram of table_list_clean[[ii]]\$pups_be



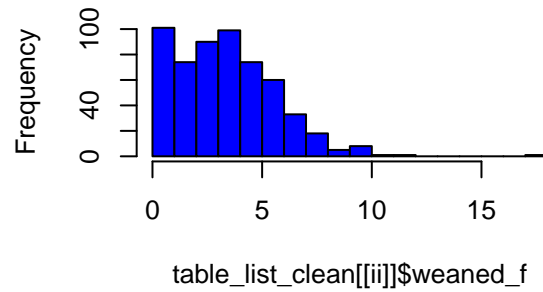
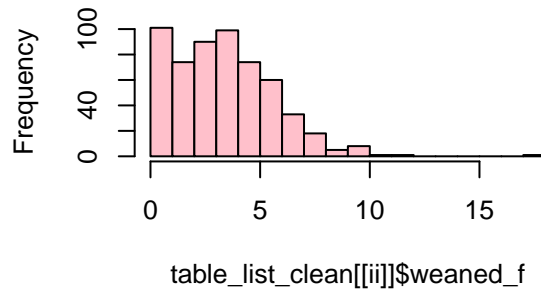
Histogram of table_list_clean[[ii]]\$weaned_f Histogram of table_list_clean[[ii]]\$weane



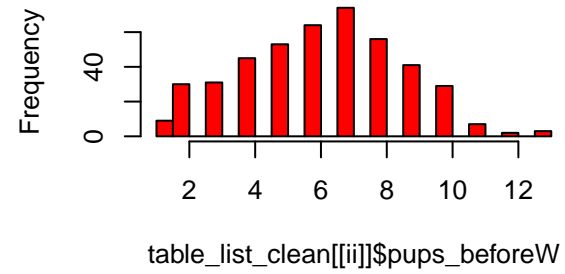
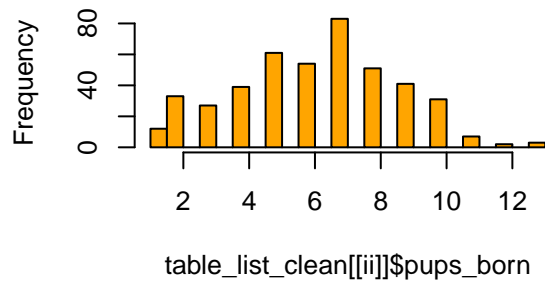
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



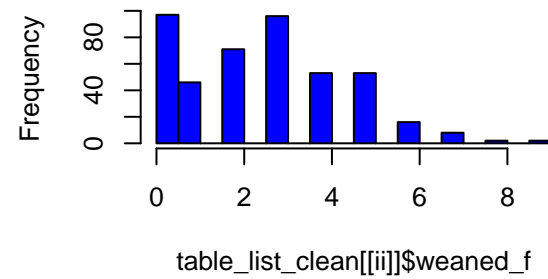
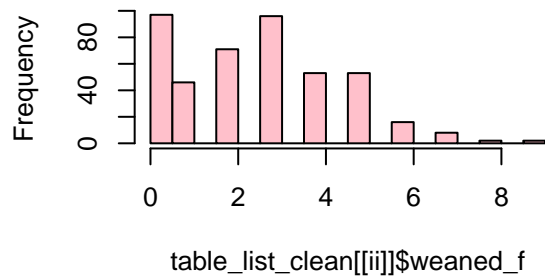
Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane



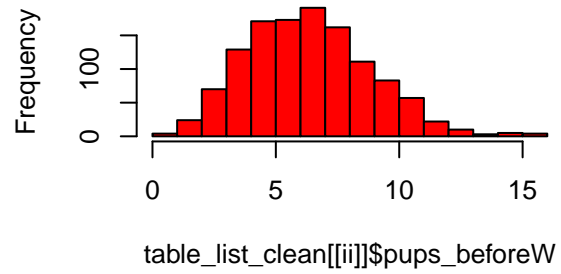
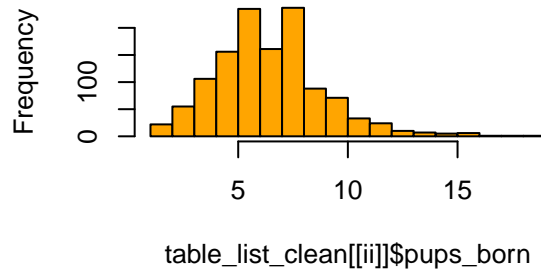
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



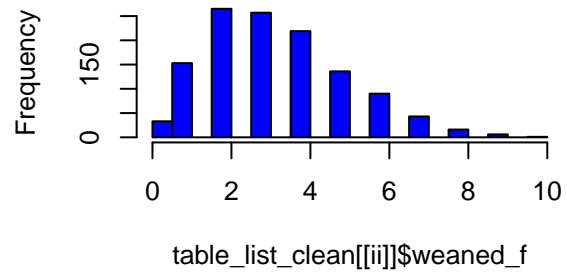
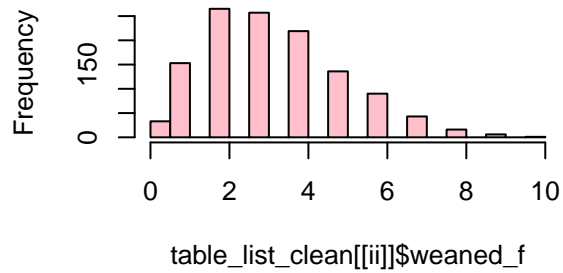
Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane



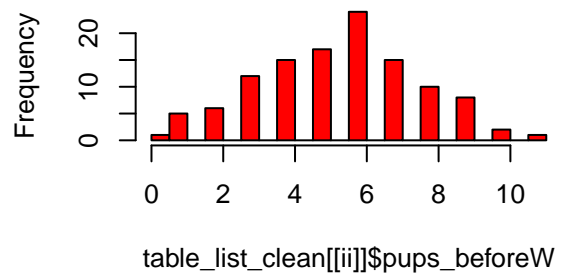
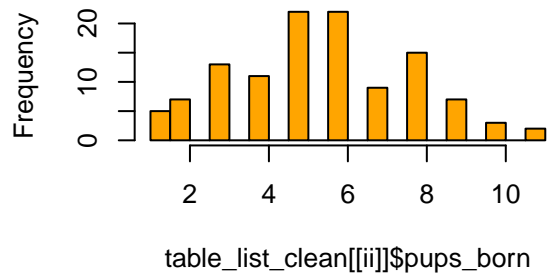
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



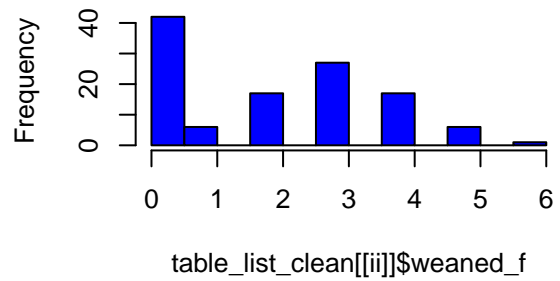
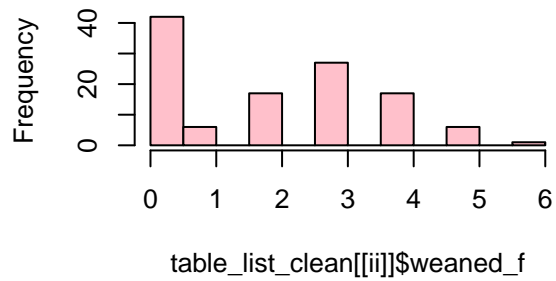
Histogram of table_list_clean[[ii]]\$weane Istogram of table_list_clean[[ii]]\$weane



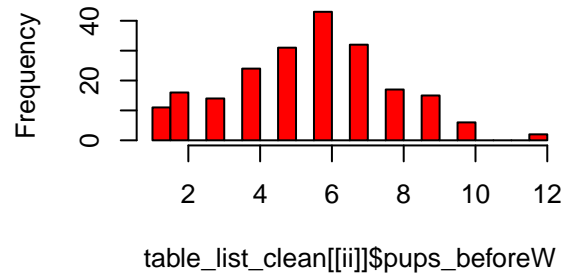
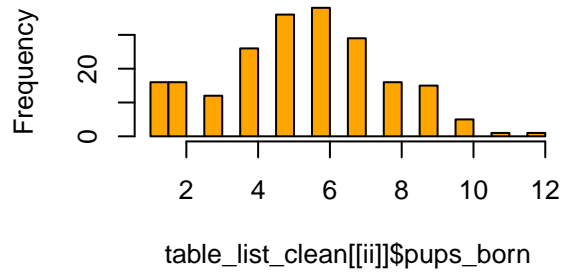
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



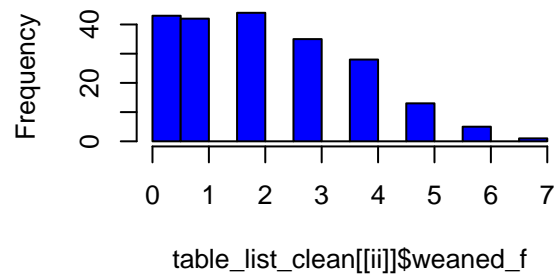
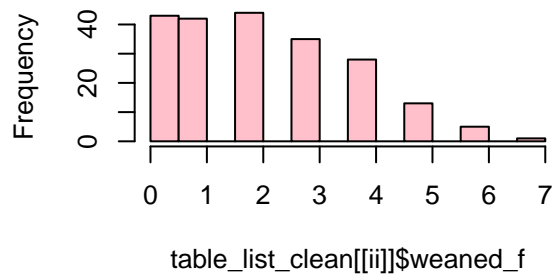
Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane



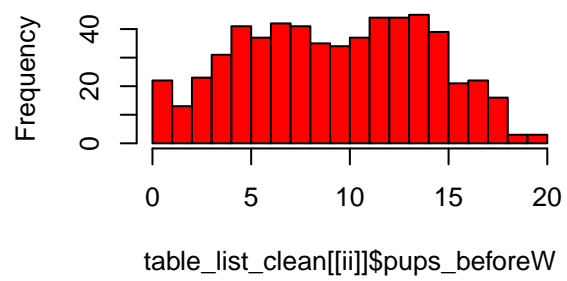
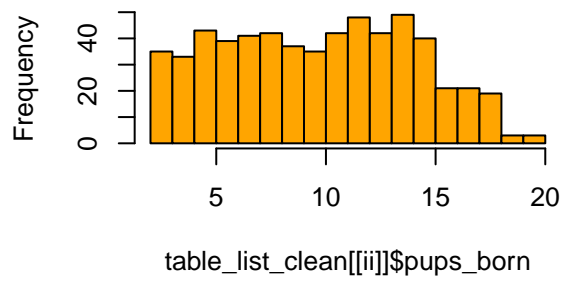
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



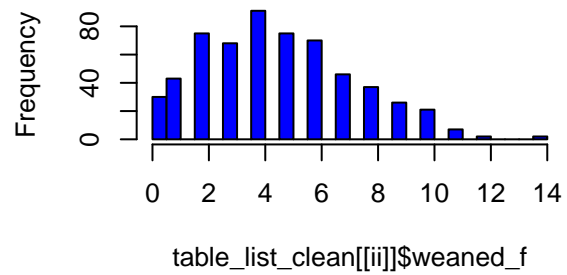
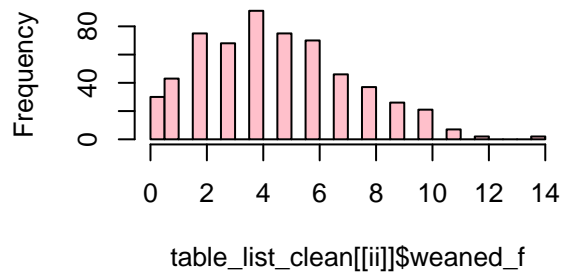
Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane



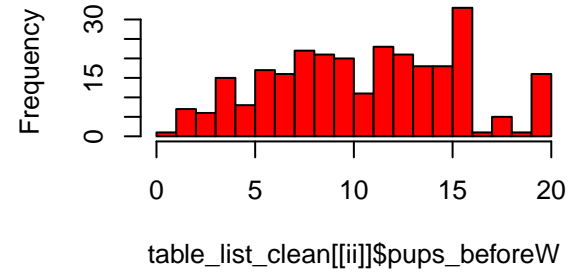
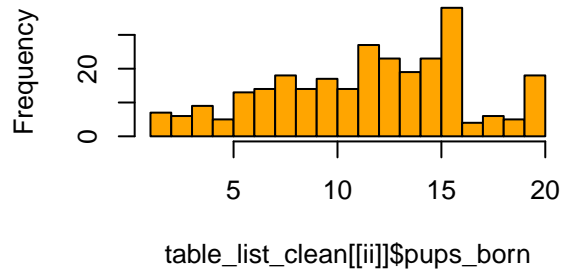
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



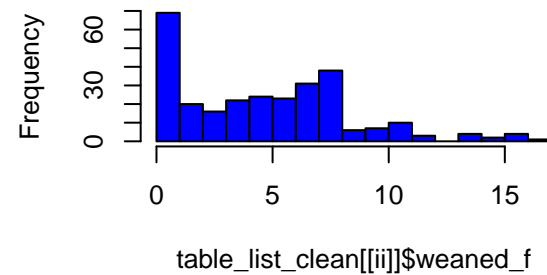
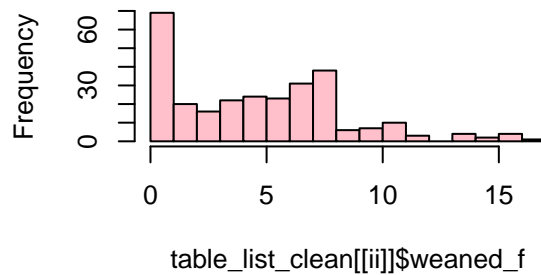
Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane



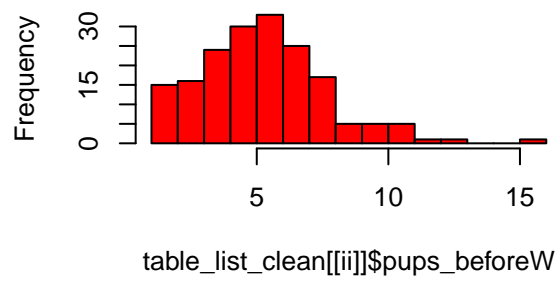
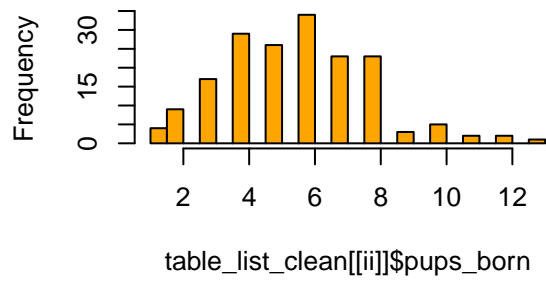
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



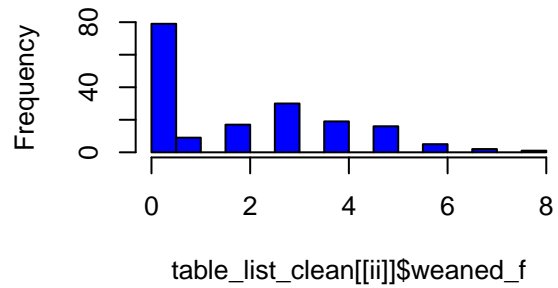
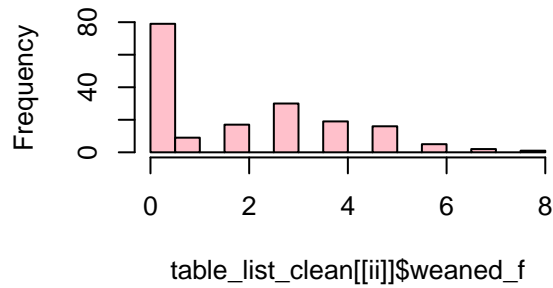
Histogram of table_list_clean[[ii]]\$weane Istogram of table_list_clean[[ii]]\$weane



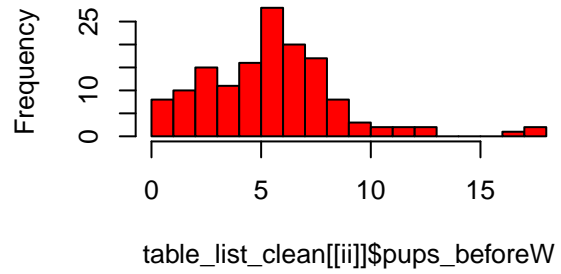
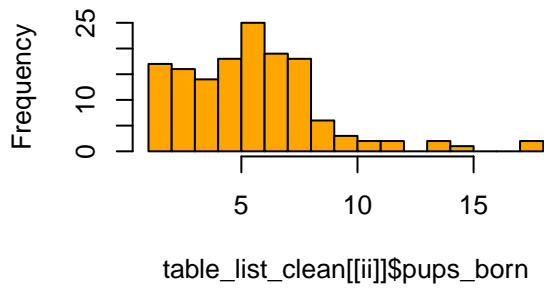
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



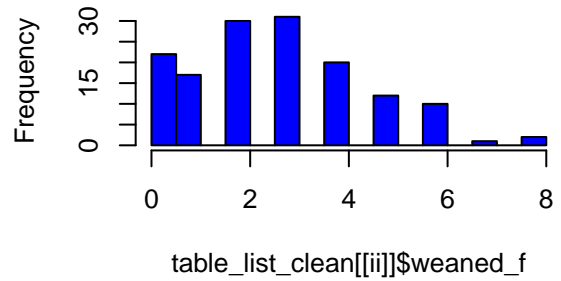
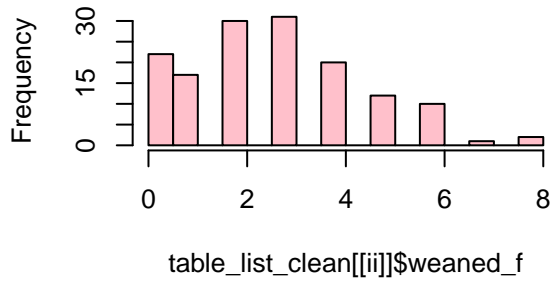
Histogram of table_list_clean[[ii]]\$weane Istogram of table_list_clean[[ii]]\$weane



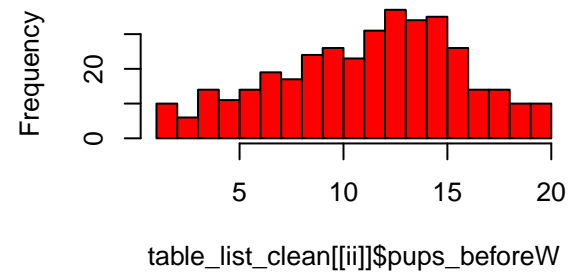
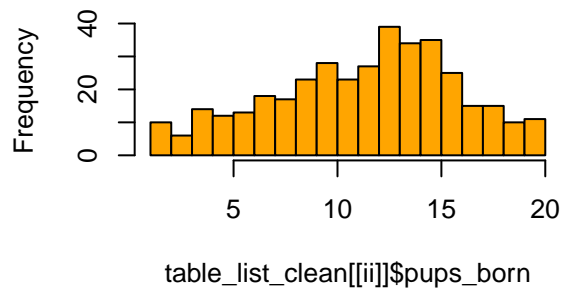
Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



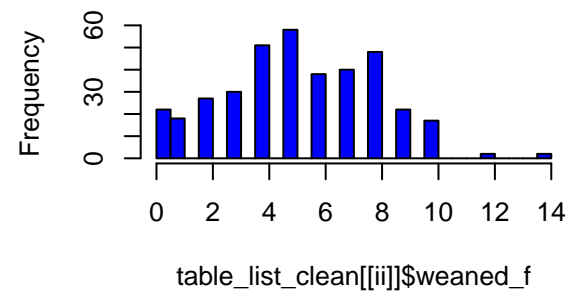
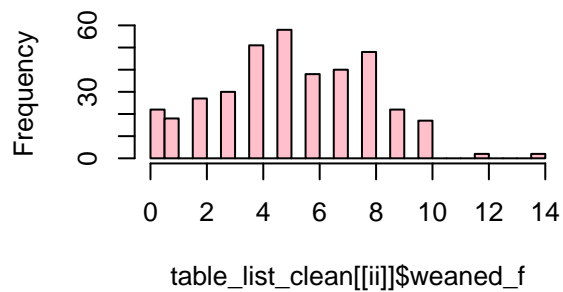
Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane



Histogram of table_list_clean[[ii]]\$pups_ Istogram of table_list_clean[[ii]]\$pups_be



Histogram of table_list_clean[[ii]]\$weane Histogram of table_list_clean[[ii]]\$weane

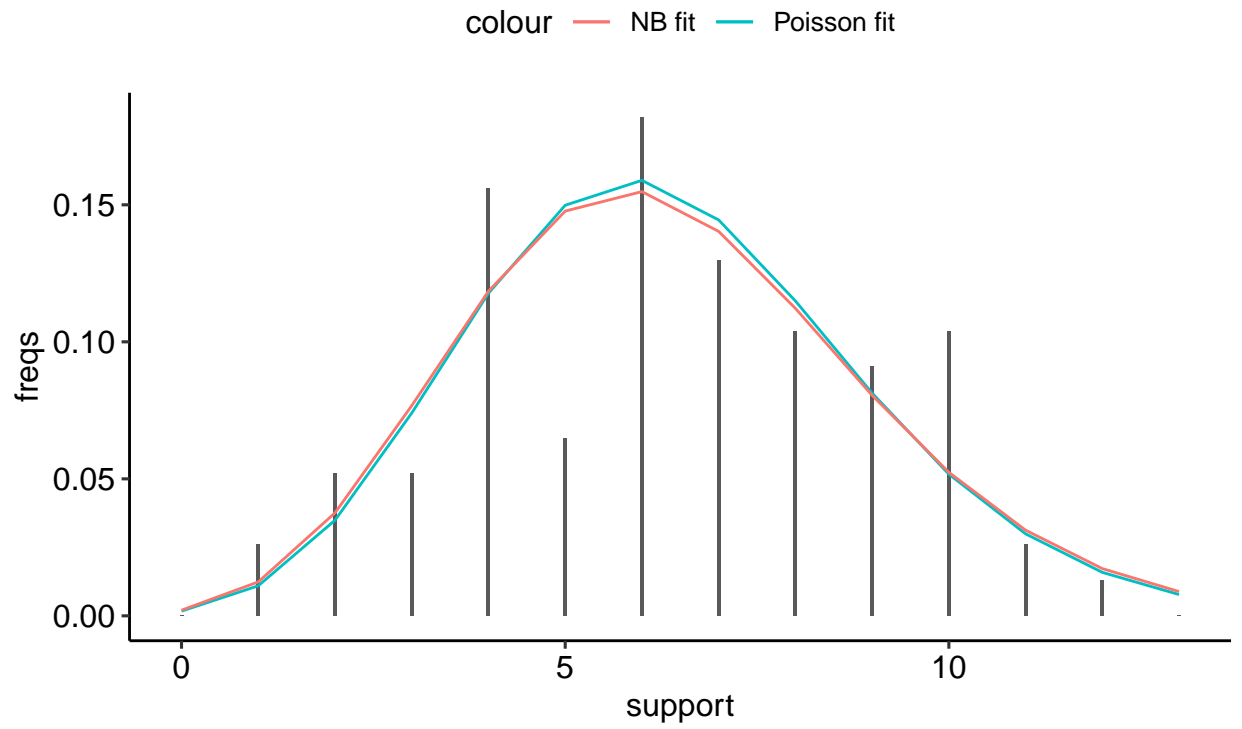


Histograms with a Pois and NB fit for each strain

```
hists_list
```

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

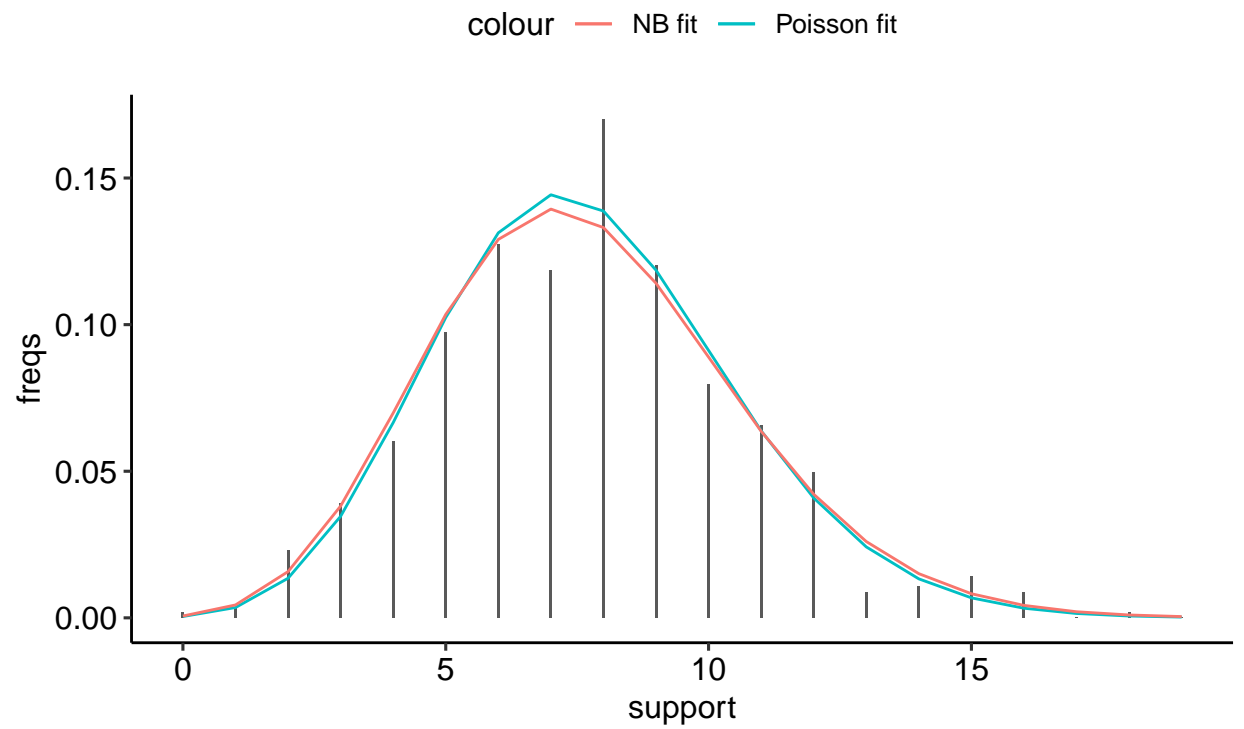
Strain: B6cBrd



Number of litters: 77

```
##  
## [[2]]
```

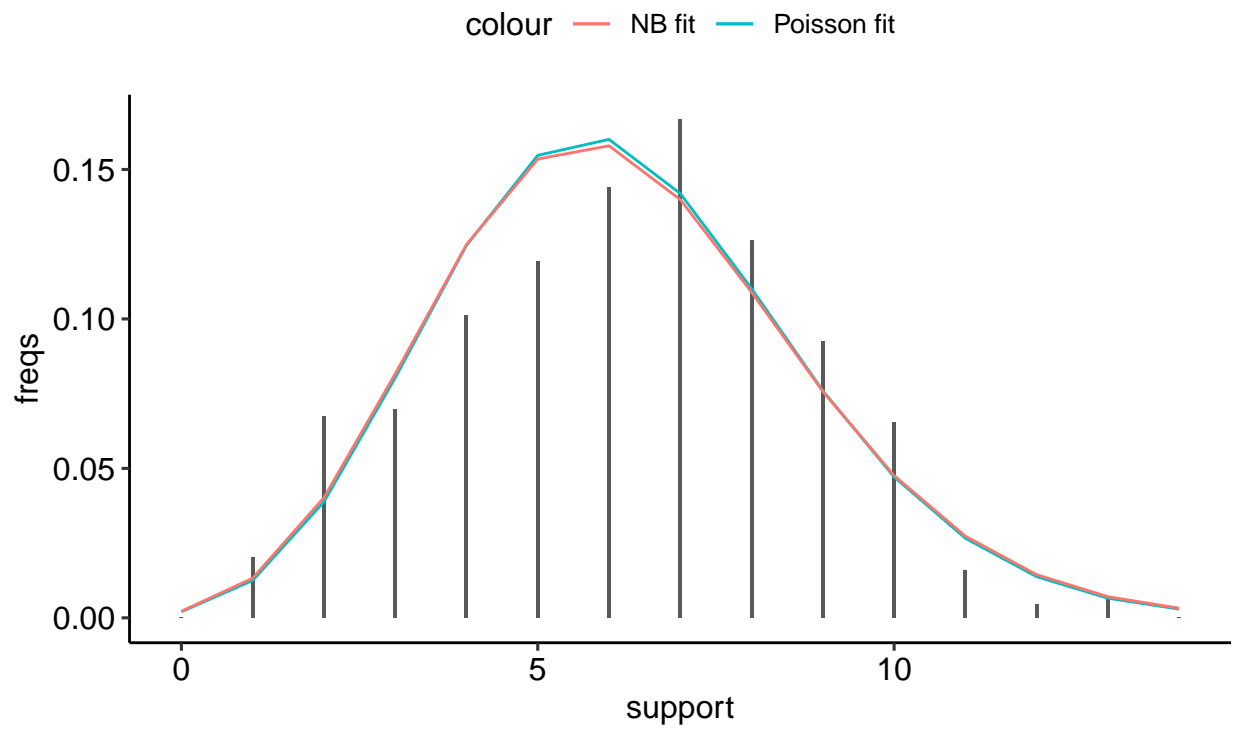

Strain: B6D2F1



Number of litters: 565

```
##  
## [[3]]
```

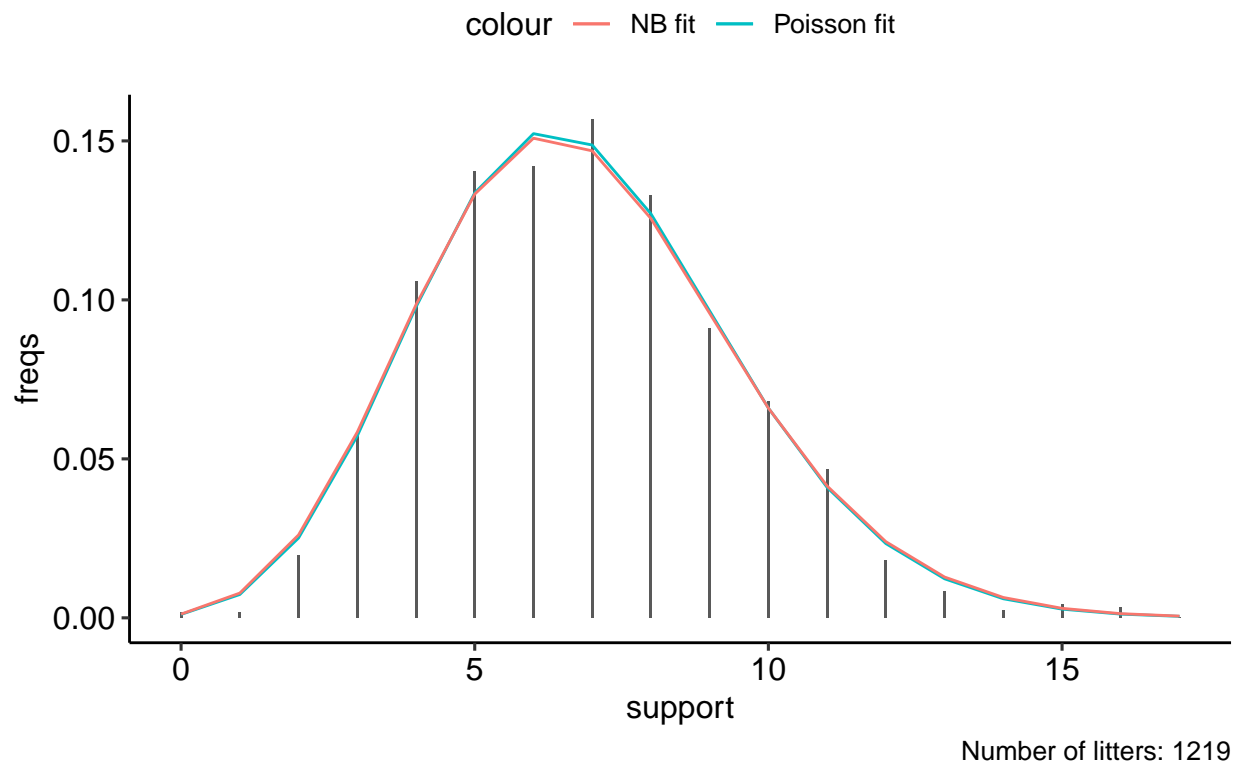
Strain: B6J_CrIF



Number of litters: 444

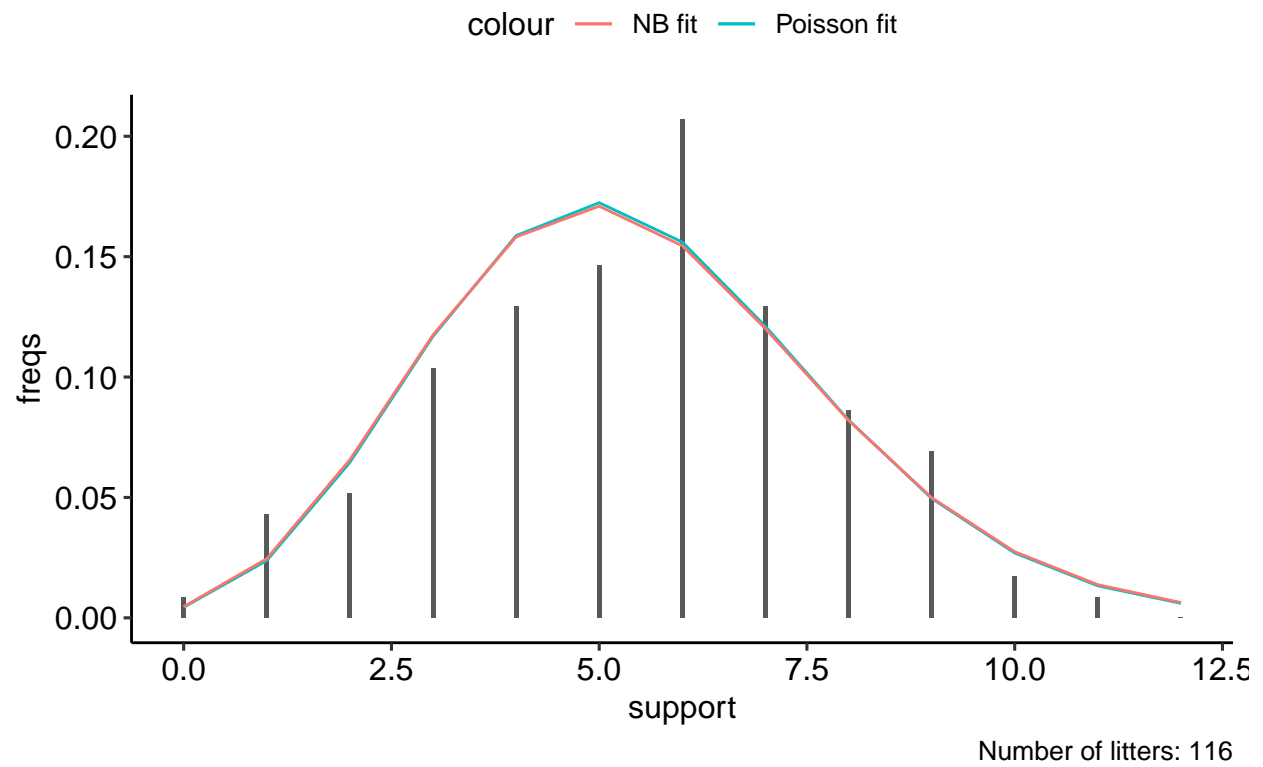
```
##  
## [[4]]
```

Strain: B6J_Fue



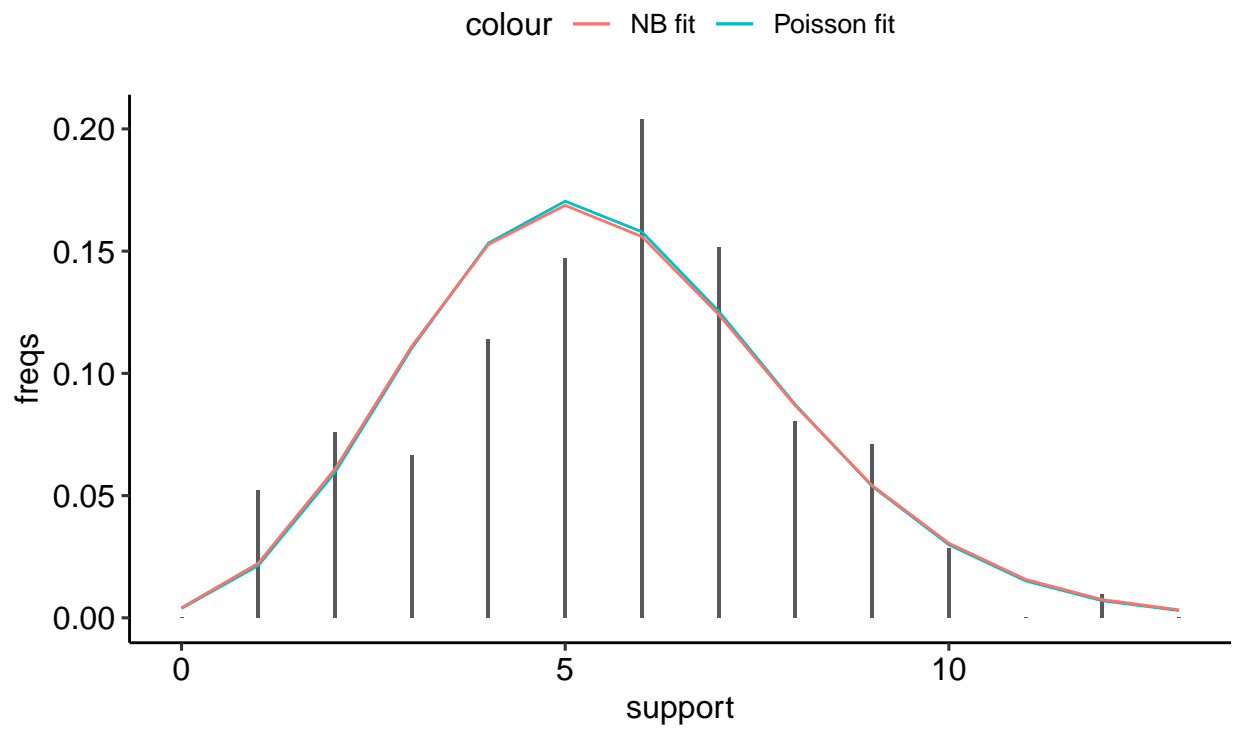
```
##  
## [[5]]
```

Strain: Balbc



```
##  
## [[6]]
```

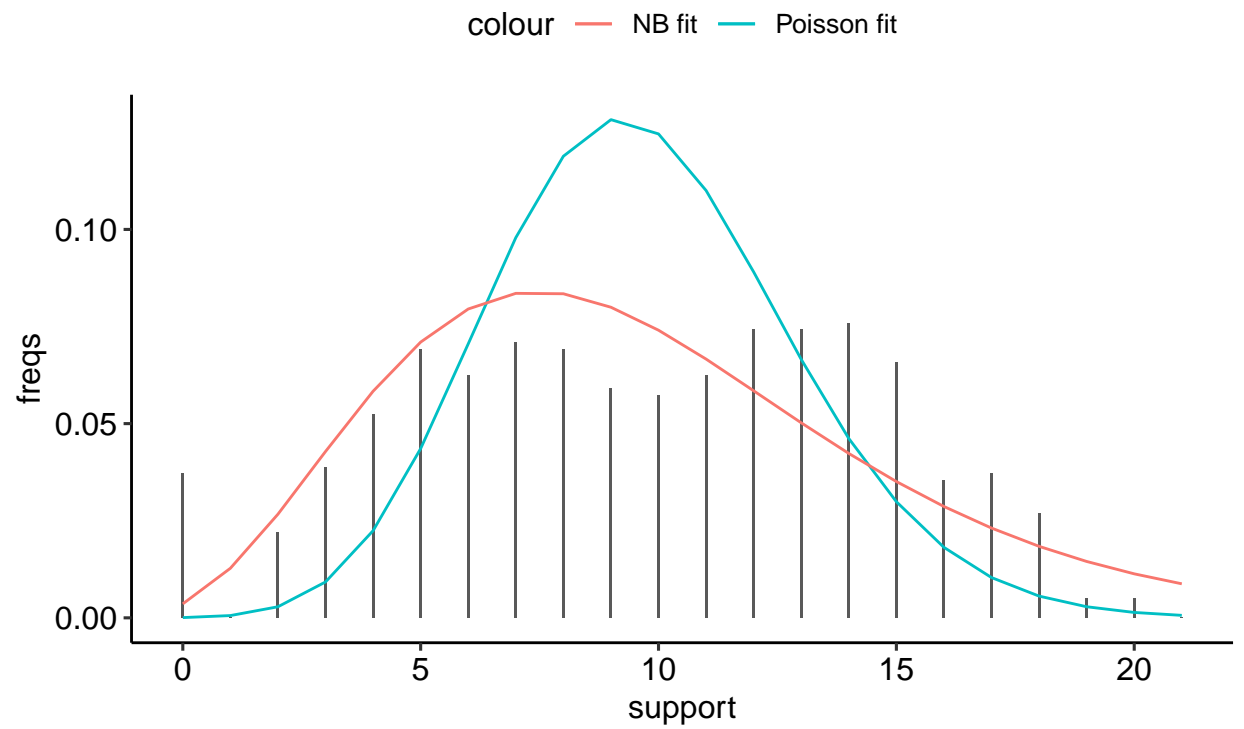
Strain: Card9_KO



Number of litters: 211

```
##  
## [[7]]
```

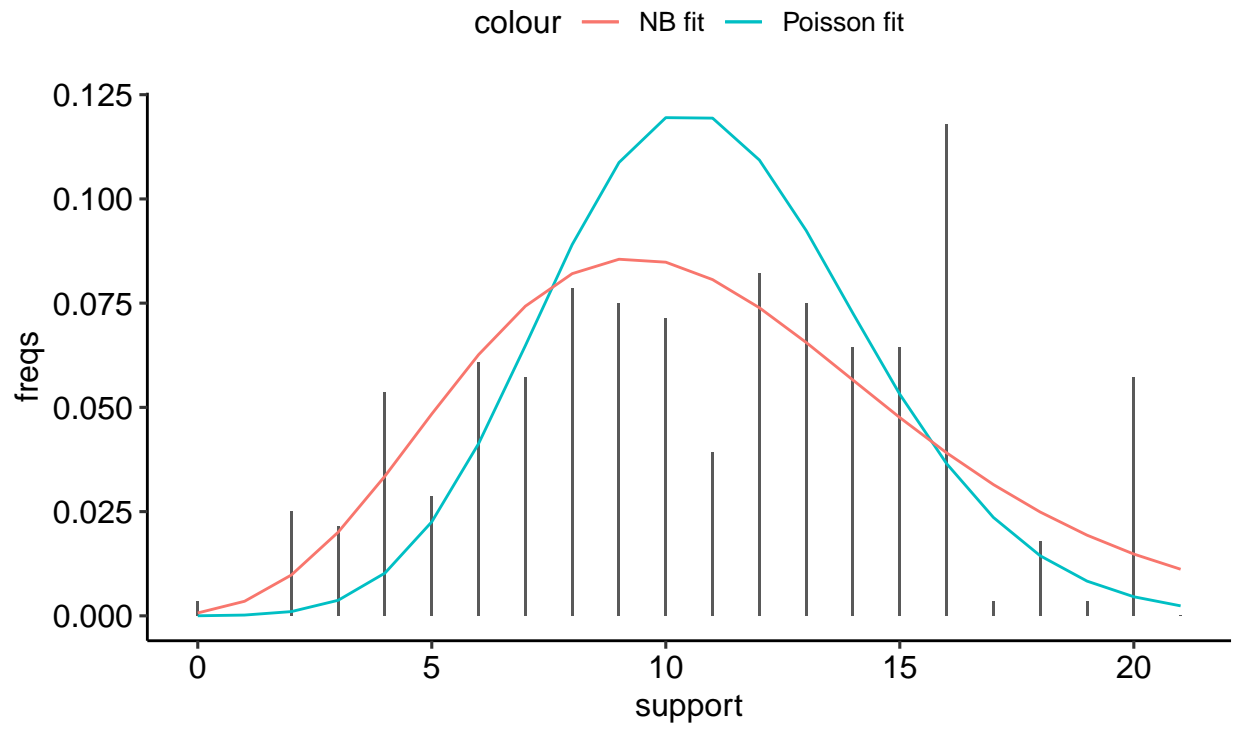
Strain: CD1_1999_2010



Number of litters: 593

```
##  
## [[8]]
```

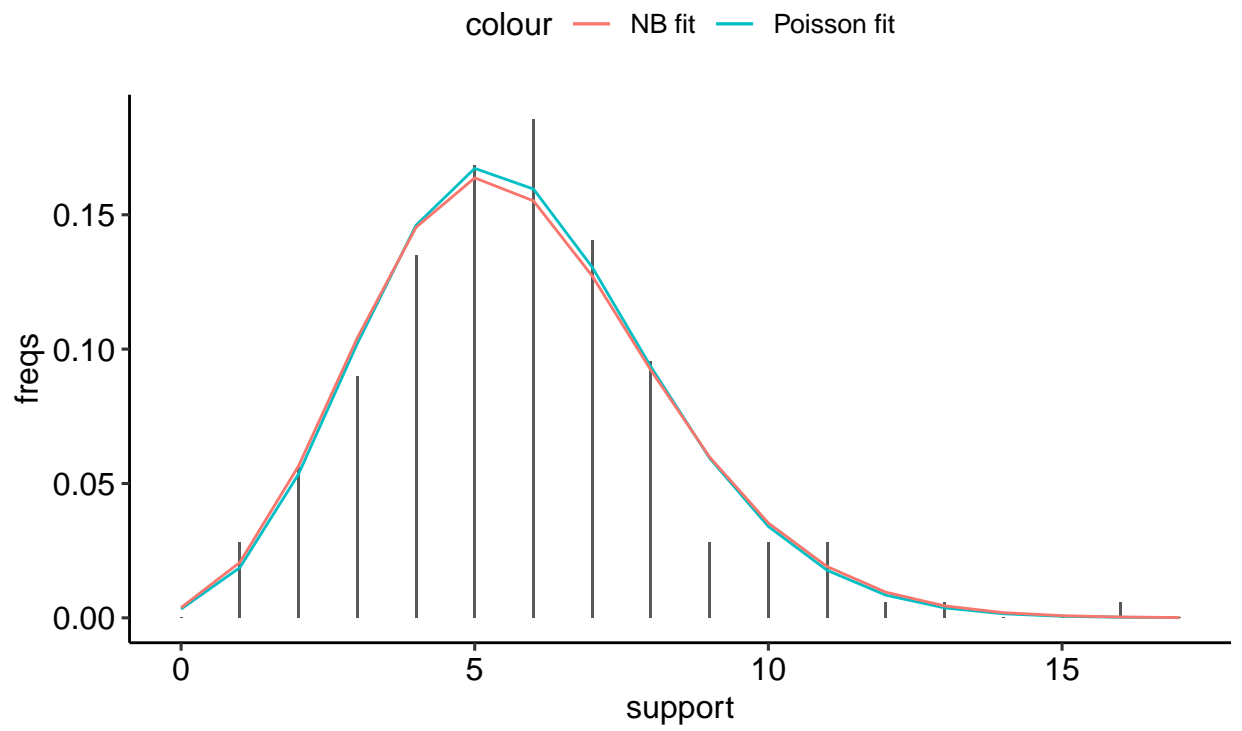
Strain: CD1_2010_2020



Number of litters: 280

```
##  
## [[9]]
```

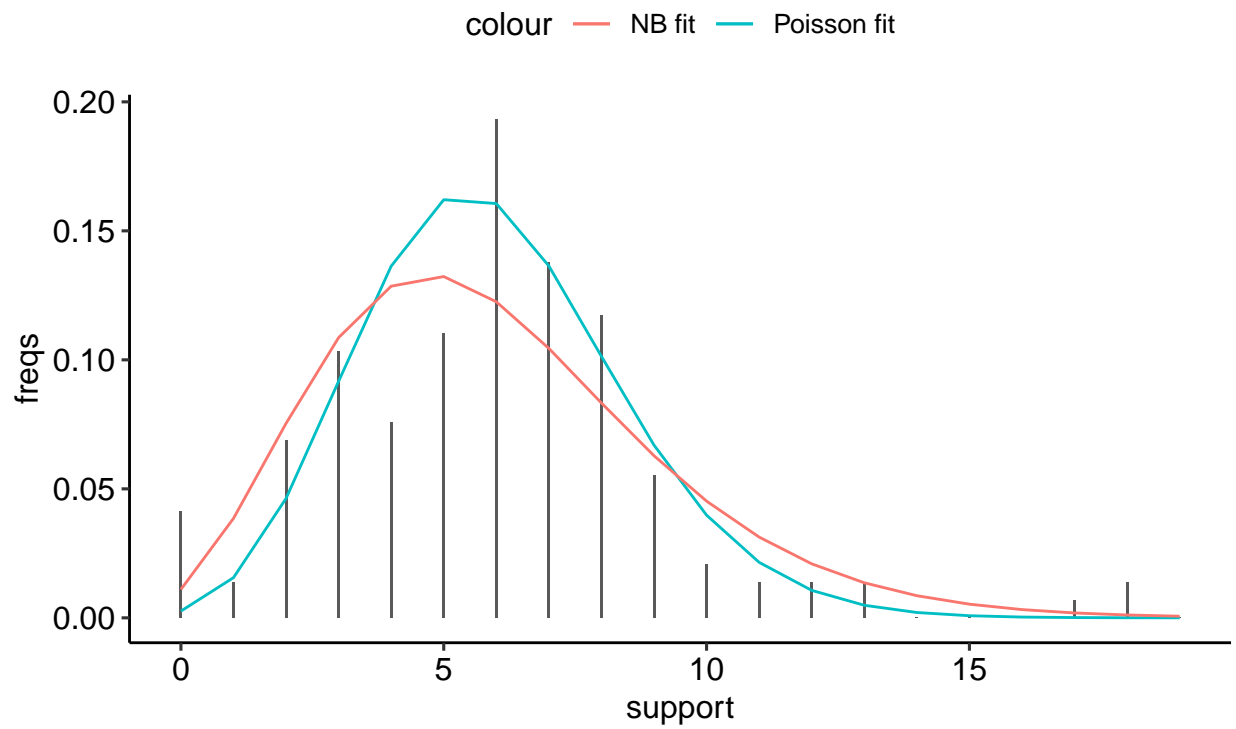
Strain: DBA2_J_Fue



Number of litters: 178

```
##  
## [[10]]
```

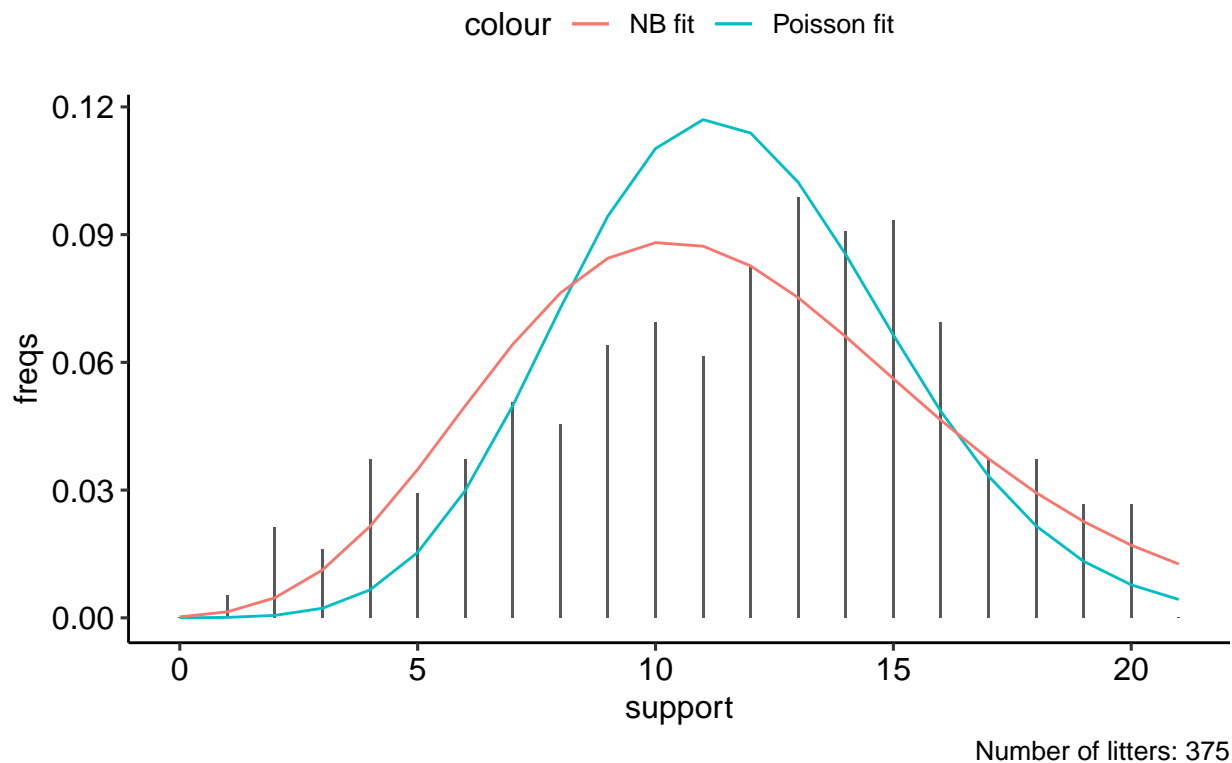

Strain: FcRn



Number of litters: 145

```
##  
## [[11]]
```

Strain: NMRI



Pretty histograms figure (Figure 4A)

We chose only those strains that look like a proper Poisson - because others are noisy.. Well, for now this is a fine criteria.

```
# "B6cBrd"      "B6D2F1"      "B6J_CrlF"      "B6J_Fue"      "Balbc"      "Card9_KO"      "CD1_
# [9] "DBA2_J_Fue"      "FcRn"      "NMRI"
table_list_merged <- list()
table_list_merged[1:2] <- table_list_clean[1:2]
table_list_merged[[3]] <- rbind(table_list_clean[["B6J_CrlF"]],table_list_clean[["B6J_Fue"]])
table_list_merged[4:8] <- table_list_clean[c("Balbc","Card9_KO","DBA2_J_Fue","FcRn","NMRI")]
table_list_merged[[9]] <- rbind(table_list_clean[["CD1_1999_2010"]],table_list_clean[["CD1_2010_2020"]])

names(table_list_merged) <- c(names(table_list_clean[1:2]),
                             "B6J",
                             c("Balbc","Card9_KO","DBA2_J_Fue","FcRn","NMRI"),
                             "CD1")

data4histo <- lapply(names(table_list_merged),
                     FUN=function(df_name){

df <- table_list_merged[[df_name]]
df_clean <- as.data.frame(
  apply(data.frame(df[,2:3]),
    c(1,2),
```

```

      as.numeric))
rowmax <- apply(df_clean, 1, max)

# can be changed to pups_born or pups_beforeW
# xx <- rowmax
xx <- df_clean$pups_beforeW
support <- seq(0, max(xx)+1)

fitNB <- MASS::fitdistr(x = xx,
                        densfun = "negative binomial")
fitPois <- MASS::fitdistr(x = xx,
                          densfun = "poisson")
fitProb_NB <- dnbinom(support,
                      size=fitNB$estimate["size"],
                      mu=fitNB$estimate["mu"])
fitProb_Pois <- dpois(support, lambda=fitPois$estimate["lambda"])

observed_values <- table(xx)
values <- sapply(
  support,
  FUN = function(z){
    ifelse(as.character(z) %in% names(observed_values),
           observed_values[as.character(z)],
           0)
  })
names(values) <- support

stopifnot(observed_values == values[names(observed_values)])

freqs <- values / sum(values)
df2plot <- data.frame(
  support,
  values,
  freqs,
  fitted_Pois = fitProb_Pois,
  fitted_NB = fitProb_NB,
  strain = rep(df_name, length(support)),
  nlitters = rep(nrow(df_clean), length(support)))

return(df2plot)
#print(freqs)
#print(barplot(height = freqs))
# ggplot(df2plot, aes(x=support)) +
#   geom_bar(aes(y=freqs), stat="identity", width = 0.05) +
#   geom_line(aes(y=fitted_Pois, color="Poisson fit")) +
#   geom_line(aes(y=fitted_NB, color="NB fit")) +
#   labs(title = paste0("Strain: ", df_name),
#        caption = paste0("Number of litters: ", nrow(df_clean)))
#
#
#
})
names(data4histo) <- names(table_list_merged)

```

```
df4histo <- do.call(
  "rbind",
  data4histo[1:7]) # excluded NMRI and CD1 aas they are not looking as Poisson
```

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse
```

```
## v tibble 3.0.3    v dplyr 1.0.2
## v tidyr  1.1.2    v stringr 1.4.0
## v readr  1.3.1    v forcats 0.5.0
## v purrr  0.3.4
```

```
## -- Conflicts ----- tidyverse_confli
```

```
## x dplyr::filter() masks stats::filter()
## x purrr::is_null() masks testthat::is_null()
## x dplyr::lag() masks stats::lag()
## x dplyr::matches() masks tidyr::matches(), testthat::matches()
```

```
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
## Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
```

```
## if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(forcats)
library(plyr)
```

```
## -----
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## -----
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
## arrange, count, desc, failwith, id, mutate, rename, summarise,
## summarize
```

```
## The following object is masked from 'package:purrr':
##
## compact
```

```
## The following object is masked from 'package:ggpubr':
##
## mutate
```

```
df4histo$strain
```

```
## [1] "B6cBrd" "B6cBrd" "B6cBrd" "B6cBrd" "B6cBrd"
## [6] "B6cBrd" "B6cBrd" "B6cBrd" "B6cBrd" "B6cBrd"
## [11] "B6cBrd" "B6cBrd" "B6cBrd" "B6cBrd" "B6D2F1"
## [16] "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1"
## [21] "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1"
## [26] "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1"
## [31] "B6D2F1" "B6D2F1" "B6D2F1" "B6D2F1" "B6J"
## [36] "B6J" "B6J" "B6J" "B6J" "B6J"
## [41] "B6J" "B6J" "B6J" "B6J" "B6J"
## [46] "B6J" "B6J" "B6J" "B6J" "B6J"
## [51] "B6J" "B6J" "Balbc" "Balbc" "Balbc"
## [56] "Balbc" "Balbc" "Balbc" "Balbc" "Balbc"
## [61] "Balbc" "Balbc" "Balbc" "Balbc" "Balbc"
## [66] "Card9_K0" "Card9_K0" "Card9_K0" "Card9_K0" "Card9_K0"
## [71] "Card9_K0" "Card9_K0" "Card9_K0" "Card9_K0" "Card9_K0"
## [76] "Card9_K0" "Card9_K0" "Card9_K0" "Card9_K0" "DBA2_J_Fue"
## [81] "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue"
## [86] "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue"
## [91] "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue" "DBA2_J_Fue"
## [96] "DBA2_J_Fue" "DBA2_J_Fue" "FcRn" "FcRn" "FcRn"
## [101] "FcRn" "FcRn" "FcRn" "FcRn" "FcRn"
## [106] "FcRn" "FcRn" "FcRn" "FcRn" "FcRn"
## [111] "FcRn" "FcRn" "FcRn" "FcRn" "FcRn"
## [116] "FcRn" "FcRn"
```

```
df4histo <- arrange(df4histo, nlitters)
# B6cBrd Balbc FcRn DBA2_J_Fue Card9_K0 B6D2F1 B6J
df4histo$strain <- plyr::revalue(df4histo$strain,
  c("B6cBrd"="B6 Albino ",
    "Balbc"="BALB/cJ",
    "FcRn"="FcRn",
    "B6J"="C57BL/6J",
    "DBA2_J_Fue"="DBA2J",
    "Card9_K0"="Card9 K0",
    "B6D2F1"="BD2F1 "
  ))
df4histo$strain <- reorder(df4histo$strain, df4histo$nlitters)

fig4a <- df4histo %>%
  ggplot(aes(x=support, color=strain, fill=strain)) +
  geom_bar(aes(y=freqs), stat = "identity", width = 0.5) +
  scale_fill_viridis(discrete=TRUE) +
  scale_color_viridis(discrete=TRUE) +
```

```

geom_line(aes(y=fitted_Pois), col = "black") +
geom_line(aes(y=fitted_NB), col="red4") +
theme_ipsum() +
theme(
  legend.position="none",
  panel.spacing = unit(1, "lines"),
  strip.text.x = element_text(size = 0)
) +
xlab("") +
ylab("") +
#labs(caption = paste0("Number of litters: ", nlitters))) +
#xlab("Number of pups in a litter") +
#ylab("Frequency") +
facet_wrap(~strain)

fig4a_withlabels <- df4histo %>%
  ggplot( aes(x=support, color=strain, fill=strain)) +
  geom_bar(aes(y=freqs), stat = "identity", width = 0.5) +
  scale_fill_viridis(discrete=TRUE) +
  scale_color_viridis(discrete=TRUE) +
  geom_line(aes(y=fitted_Pois), col = "black") +
  geom_line(aes(y=fitted_NB), col="red4") +
  theme_ipsum() +
  theme(
    legend.position="none",
    panel.spacing = unit(1, "lines"),
    strip.text.x = element_text(size = 10)
  ) +
  xlab("Number of pups in a litter; NB fit: red, Poison fit: black") +
  ylab("Frequency") +
  facet_wrap(~strain*nlitters)

pdf("./figures/fig4A/fig4a_nolabels.pdf")
print(fig4a)
dev.off()

pdf("./figures/fig4A/fig4a_withlabels.pdf")
print(fig4a_withlabels)
dev.off()

```

Litter means etc. by strain

```
names(table_list_clean)
```

```
## [1] "B6cBrd"      "B6D2F1"      "B6J_Cr1F"    "B6J_Fue"
## [5] "Balbc"       "Card9_K0"    "CD1_1999_2010" "CD1_2010_2020"
## [9] "DBA2_J_Fue"  "FcRn"        "NMRI"
```

```

mousebreeding_by_strain_clean <- c(
  table_list_clean[c('B6cBrd', 'B6D2F1', 'B6J_Cr1F', 'B6J_Fue',
                    'Balbc',
                    'Card9_KO', 'DBA2_J_Fue', 'FcRn', 'NMRI')],
  list(rbind(table_list_clean[["CD1_1999_2010"]],
             table_list_clean[["CD1_2010_2020"]]))
)
names(mousebreeding_by_strain_clean)[length(mousebreeding_by_strain_clean)] <- "CD1"

```

Fertility

```

load("../external/mouse_data/mouse3/fertility/nfemales.Rdata")
pre_tab_list <- lapply(
  names(table_list_clean), FUN = function(df_name){
    df <- table_list_clean[[df_name]]
    originalPDF_file <- summaryDF$originalPDF[which(
      summaryDF$strainNameShort == df_name)]
    df_clean <- as.data.frame(
      apply(data.frame(df[,2:3]),
            c(1,2),
            as.numeric))
    rowmax <- apply(df_clean, 1, max)
    c(
      round(mean(rowmax),1),
      median(rowmax),
      max(rowmax),
      length(rowmax),
      nrow(mousebreeding[[df_name]]),
      nfemales[originalPDF_file],
      df_name)
  })

```

```
sum(mousebreeding[[4]]$pups_born==0)
```

```
## [1] 0
```

```

pre_tab_df <- data.frame(do.call(rbind, pre_tab_list))
colnames(pre_tab_df) <- c("mean", "median", "max", "n_trust_breedings", "n_tot_breedings", "n_fertile",

#fertility <- round(
#  as.numeric(pre_tab_df$n_fertile)/as.numeric(pre_tab_df$n_tot_breedings)*100,
#  digits = 1)

strain_stats <- data.frame(
  strain = pre_tab_df[,c("strain")],
  #fertility = fertility,
  litter_mean = as.numeric(pre_tab_df[,c("mean")]),
  litter_median = as.numeric(pre_tab_df[,c("median")]),
  litter_max = as.numeric(pre_tab_df[,c("max")]),
  n_trust_breedings = as.numeric(pre_tab_df[,c("n_trust_breedings")]),

```

```
    n_tot_breedings = as.numeric(pre_tab_df[,c("n_tot_breedings")])
    #n_fertile = as.numeric(pre_tab_df[,c("n_fertile")])
  )

save(strain_stats, file = "strain_stats.Rdata")
write.table(strain_stats, file = "strain_stats.txt",
            quote = FALSE, sep = "\t",
            row.names = F, col.names = T)
```