MBP intro to statistics bootcamp

https://jasonlerch.github.io/MBP-stats-2019/

Jason Lerch

Day 1

Hello World

The three challenges of statistical inference are[1]:

1. Generalizing from sample to population
2. Generalizing from control to treatment group
3. Generalizing from observed measurements to underlying constructs of interest

[1] From Andrew Gelman

Three laws of statistics

Arthur C. Clarke's three laws[1]:

1. When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.
2. The only way of discovering the limits of the possible is to venture a little way past them into the impossible.
3. Any sufficiently advanced technology is indistinguishable from magic.

Andrew Gelman's updates[2]:

1. When a distinguished but elderly scientist states that “You have no choice but to accept that the major conclusions of these studies are true,” don’t believe him.
2. The only way of discovering the limits of the reasonable is to venture a little way past them into the unreasonable.
3. Any sufficiently crappy research is indistinguishable from fraud.

[1] https://en.wikipedia.org/wiki/Clarke%27s_three_laws

[2] http://andrewgelman.com/2016/06/20/clarkes-law-of-research/

The MBP statistics bootcamp

Goals of this week:

1. Teach the theory and practice of statistics
2. Applied data analysis problem solving using R
3. Think hard about truth and replicability in science

Slides, recommended readings, and extra resources here:

https://jasonlerch.github.io/MBP-stats-2019/

(Will try and have slides for each day up the night before)

The MBP statistics bootcamp

Hour	Monday	Tuesday	Wednesday	Thursday	Friday
9-12	Introduction. Data organization, descriptive statistics, plotting, basic models.	Probability in all its glory. Multiple linear models, interactions, p values.	Hypothesis testing, searching for truth, multiple comparisons, and the crisis of replicability	Putting it all together – analyzing a biomedical dataset from beginning to end. Review	Presentations, exam
12-3	Group assignment #1	Group assignment #2	Group assignment #3	Group assignment #4	

Grading

Exams (concepts only, no R):

What	When	How much
Short exam	Tuesday	5%
Short exam	Wednesday	5%
Short exam	Thursday	5%
Final exam	Friday	35%

Group assignments and presentations (R analyses and concepts):

What	Due when	How much
Group assignment #1	Tuesday	10%
Group assignment #2	Wednesday	10%
Group assignment #3	Thursday	10%
Group assignment #4	Friday	10%

Exams

- true/false, multiple choice, and short paragraphs.
- each class begins with ~ 10 minute, short exam covering previous day.
- final exam 30-60 minutes.

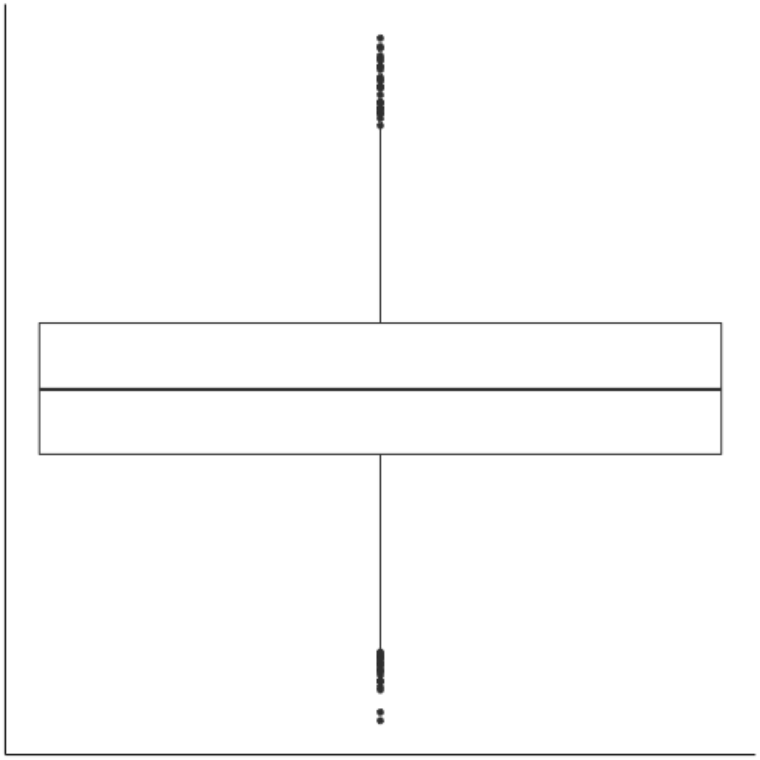
Sample questions:

Describe the null hypothesis

Identify elements of a box and whiskers plot (on a drawing)

Discuss analysis pre-registration advantages and disadvantages

TRUE/FALSE: if you compute a 95% confidence interval, you have a 95% chance of it containing the true value



Group assignments

- split into small groups of 3-4.
- we will assign groups.
- will try to mix groups by R and programming expertise.
- each group will be graded as a unit.
- final presentation given by a member of the group with least R/programming expertise.

Let's get started

Statistical software

Common software

1. Excel
2. SPSS
3. SAS
4. matlab
5. python
6. R

Ups and downs of R

1. Open source, free, and powerful.
2. If a statistical test exists, it likely exists in R.
3. Literate programming/self documenting analyses.
4. Very strong in bioinformatics.
5. Steeper learning curve.

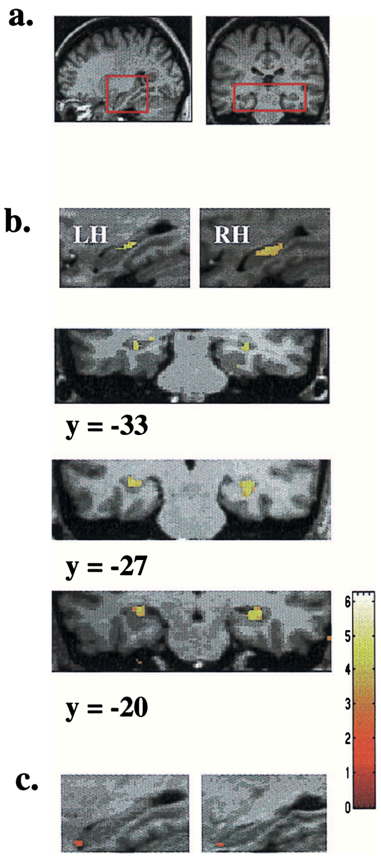
Reading and summarizing our data

Intro to our dataset

How do our brains change as we learn or undergo new experiences?

Earliest evidence that our brains are *plastic* at larger, or *mesoscopic*, scales came from a study of taxi drivers in London, UK.

Mechanism of how that happens is unclear.



Mouse models

We can create taxi driving mice.

Use high-field MRI to get similar readout as in humans.

Use genetic models to test hypotheses of implicated pathways.

Use RNA sequencing to assess what changes per genotype or experimental group.

The dataset

There are 283 mice in this dataset, with MRI scans acquired at 6 timepoints.

We have 3 genotypes: CREB -/-, CREB +/-, CREB +/+

There are 4 environmental conditions: Enriched, Exercise, Isolated Standard, Standard

MRIs were acquired at every timepoint, and the brains automatically segmented into regions.

There are good reasons to believe that the hippocampus and the dentate gyrus of the hippocampus will be the most affected by the environmental interventions.

The effect of the three genotypes alone is interesting.

Enrichment

Reading data

A surprising amount of time in data analysis is spent in prepping data for visualization and analysis.

```r
library(tidyverse)
library(forcats)

mice <- read_csv("mice.csv")
```

```
## Parsed with column specification:
## cols(
##   Age = col_double(),
##   Sex = col_character(),
##   Condition = col_character(),
##   Mouse.Genotyping = col_character(),
##   ID = col_double(),
##   Timepoint = col_character(),
##   Genotype = col_character(),
##   DaysOfEE = col_double(),
##   DaysOfEE0 = col_double()
## )
```

Meet the mice

```
str(mice, give.attr=FALSE)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':    1392 obs. of  9 v
##  $ Age             : num  8.5 8.5 8.5 9.5 9.5 8.5 8.5 9.5 8.5 9.5 ...
##  $ Sex             : chr  "M" "M" "M" "M" ...
##  $ Condition       : chr  "Enriched" "Standard" "Standard" "Enriched" ...
##  $ Mouse.Genotyping: chr  "Heterozygous" "Heterozygous" "Heterozygous" "Wildt
##  $ ID              : num  901 899 898 891 893 901 899 889 898 895 ...
##  $ Timepoint       : chr  "Pre1" "Pre1" "Pre1" "Pre1" ...
##  $ Genotype        : chr  "CREB +/-" "CREB +/-" "CREB +/-" "CREB +/+" ...
##  $ DaysOfEE        : num  -4 -4 -4 -4 -4 -3 -3 -3 -3 -3 ...
##  $ DaysOfEE0       : num  0 0 0 0 0 0 0 0 0 0 ...
```

Numeric variable: age

```
mice %>%
  summarise(mean=mean(Age),
            min=min(Age),
            max=max(Age))
```

mean	min	max
6.58	3.1	10.1

Factors: Sex, Condition, Genotype

```
mice %>%
  group_by(Sex) %>%
  summarise(n=n())
```

Sex	n
F	543
M	849

```
mice %>%
  group_by(Genotype) %>%
  summarize(n=n())
```

Genotype	n
CREB -/-	426
CREB +/-	486
CREB +/+	480

Subject descriptors: ID and Timepoint

```
mice %>%
  select(ID, Timepoint) %>%
  head
```

ID	Timepoint
901	Pre1
899	Pre1
898	Pre1
891	Pre1
893	Pre1
901	Pre2

Alternate encodings: Genotype

```
mice %>%
  select(Genotype, Mouse.Genotyping) %>%
  head
```

Genotype	Mouse.Genotyping
CREB +/-	Heterozygous
CREB +/-	Heterozygous
CREB +/-	Heterozygous
CREB +/+	Wildtype
CREB +/+	Wildtype
CREB +/-	Heterozygous

Alternate encodings: Days of EE, DaysofEE0

```
mice %>%
  filter(ID == 901) %>%
  select(Timepoint, DaysOfEE, DaysOfEE0) %>%
  head
```

Timepoint	DaysOfEE	DaysOfEE0
Pre1	-4	0
Pre2	-3	0
24h	1	1
48h	2	2
1 week	8	8
2 week	16	16

Overview of subject numbers

```
with(mice,
     ftable(Condition, Genotype, Timepoint))
```

```
##                            Timepoint 1 week 2 week 24h 48h Pre1 Pre2
## Condition         Genotype                                          
## Enriched          CREB -/-                 24     25   7  24   24   24
##                   CREB +/-                 30     33  12  34   30   33
##                   CREB +/+                 27     30   8  30   27   28
## Exercise          CREB -/-                 22     21   0  21   20   18
##                   CREB +/-                 17     18   0  18   17   15
##                   CREB +/+                 19     19   0  19   19   16
## Isolated Standard CREB -/-                 14     14   0  14   13   14
##                   CREB +/-                 12     12   0  12   11   12
##                   CREB +/+                 17     17   0  17   17   14
## Standard          CREB -/-                 23     26   4  26   25   23
##                   CREB +/-                 29     34   9  34   32   32
##                   CREB +/+                 28     31   6  31   31   29
```

Factors, revisited

The Timepoint order makes no sense. Let's reorder

```
mice <- mice %>%
  mutate(Timepoint=fct_relevel(Timepoint, "Pre1", "Pre2", "24h",
                               "48h", "1 week", "2 week"))
with(mice, ftable(Condition, Genotype, Timepoint))
```

```
##                            Timepoint Pre1 Pre2 24h 48h 1 week 2 week
## Condition         Genotype                                          
## Enriched          CREB -/-             24   24   7  24     24     25
##                   CREB +/-             30   33  12  34     30     33
##                   CREB +/+             27   28   8  30     27     30
## Exercise          CREB -/-             20   18   0  21     22     21
##                   CREB +/-             17   15   0  18     17     18
##                   CREB +/+             19   16   0  19     19     19
## Isolated Standard CREB -/-             13   14   0  14     14     14
##                   CREB +/-             11   12   0  12     12     12
##                   CREB +/+             17   14   0  17     17     17
## Standard          CREB -/-             25   23   4  26     23     26
##                   CREB +/-             32   32   9  34     29     34
##                   CREB +/+             31   29   6  31     28     31
```

Redo in tidyverse

```
mice %>%
  group_by(Condition, Genotype, Timepoint) %>%
  summarise(n=n()) %>% spread(Timepoint, value=n)
```

```
## # A tibble: 12 x 8
## # Groups:   Condition, Genotype [12]
##    Condition         Genotype  Pre1  Pre2 `24h` `48h` `1 week` `2 week`
##    <chr>             <chr>    <int> <int> <int> <int>    <int>    <int>
##  1 Enriched          CREB -/-    24    24     7    24       24       25
##  2 Enriched          CREB +/-    30    33    12    34       30       33
##  3 Enriched          CREB +/+    27    28     8    30       27       30
##  4 Exercise          CREB -/-    20    18    NA    21       22       21
##  5 Exercise          CREB +/-    17    15    NA    18       17       18
##  6 Exercise          CREB +/+    19    16    NA    19       19       19
##  7 Isolated Standard CREB -/-    13    14    NA    14       14       14
##  8 Isolated Standard CREB +/-    11    12    NA    12       12       12
##  9 Isolated Standard CREB +/+    17    14    NA    17       17       17
## 10 Standard          CREB -/-    25    23     4    26       23       26
## 11 Standard          CREB +/-    32    32     9    34       29       34
## 12 Standard          CREB +/+    31    29     6    31       28       31
```

Reading more data

```
volumes <- read_csv("volumes.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Timepoint = col_character()
## )

## See spec(...) for full column specifications.
```

Inspecting the new data

```
str(volumes)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':     1392 obs. of  161
##  $ amygdala                                           : num  9.84 10.3 10.53 1
##  $ anterior commissure: pars anterior                 : num  1.42 1.48 1.5 1.4
##  $ anterior commissure: pars posterior                : num  0.392 0.428 0.382
##  $ basal forebrain                                    : num  4.72 4.96 4.93 5.
##  $ bed nucleus of stria terminalis                    : num  1.24 1.31 1.28 1.
##  $ cerebellar peduncle: inferior                      : num  0.908 0.967 0.919
##  $ cerebellar peduncle: middle                        : num  1.23 1.31 1.26 1.
##  $ cerebellar peduncle: superior                      : num  0.991 0.848 0.905
##  $ cerebral aqueduct                                  : num  0.373 0.44 0.42 0
##  $ cerebral peduncle                                  : num  2.58 2.54 2.6 2.5
##  $ colliculus: inferior                               : num  5.46 5.6 5.34 5.6
##  $ colliculus: superior                               : num  9.52 9.83 9.37 9.
##  $ corpus callosum                                    : num  14.1 14.3 14 14.6
##  $ corticospinal tract/pyramids                       : num  1.59 1.59 1.56 1.
##  $ cuneate nucleus                                    : num  0.27 0.254 0.286
##  $ dentate gyrus of hippocampus                       : num  3.96 3.92 3.95 4.
##  $ facial nerve (cranial nerve 7)                     : num  0.221 0.233 0.256
##  $ fasciculus retroflexus                             : num  0.214 0.212 0.192
##  $ fimbria                                            : num  2.67 2.9 2.73 2.7
```

Linking data

```
volumes %>%
  select(ID, Timepoint) %>%
  head
```

ID	Timepoint
901	Pre1
899	Pre1
898	Pre1
891	Pre1
893	Pre1
901	Pre2

```
mice %>%
  select(ID, Timepoint) %>%
  head
```

ID	Timepoint
901	Pre1
899	Pre1
898	Pre1
891	Pre1
893	Pre1
901	Pre2

Joining data

```
mice <- mice %>%
  inner_join(volumes)
```

```
## Joining, by = c("ID", "Timepoint")

## Warning: Column `Timepoint` joining factor and character vector, coercing
## into character vector
```

```
str(mice)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':    1392 obs. of  168
##  $ Age                                              : num  8.5 8.5 8.5 9.5 9
##  $ Sex                                              : chr  "M" "M" "M" "M" .
##  $ Condition                                        : chr  "Enriched" "Stand
##  $ Mouse.Genotyping                                 : chr  "Heterozygous" "H
##  $ ID                                               : num  901 899 898 891 8
##  $ Timepoint                                        : chr  "Pre1" "Pre1" "Pr
##  $ Genotype                                         : chr  "CREB +/-" "CREB
##  $ DaysOfEE                                         : num  -4 -4 -4 -4 -4 -3
##  $ DaysOfEE0                                        : num  0 0 0 0 0 0 0 0 0
##  $ amygdala                                         : num  9.84 10.3 10.53 1
##  $ anterior commissure: pars anterior               : num  1.42 1.48 1.5 1.4
```

Data visualization

Data visualization

Data visualization communicates your data to your audience - and can be how your data communicates with you.

Excellent guide to visualization:

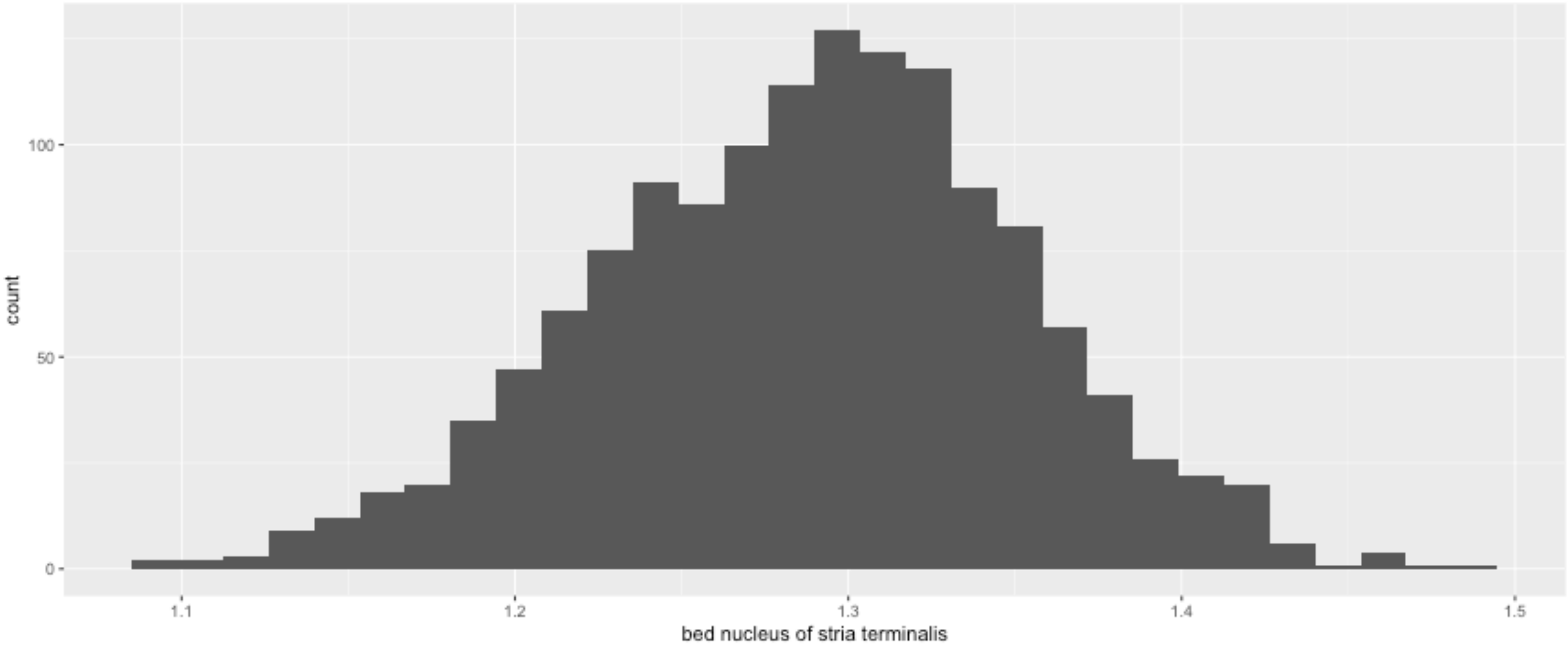
https://www.data-to-viz.com

Your task for later will be to look at the interesting variables in this dataset. For now, we will look at sex and the brain instead.

Histogram

```
ggplot(mice) +
  aes(x=`bed nucleus of stria terminalis`) +
  geom_histogram()
```

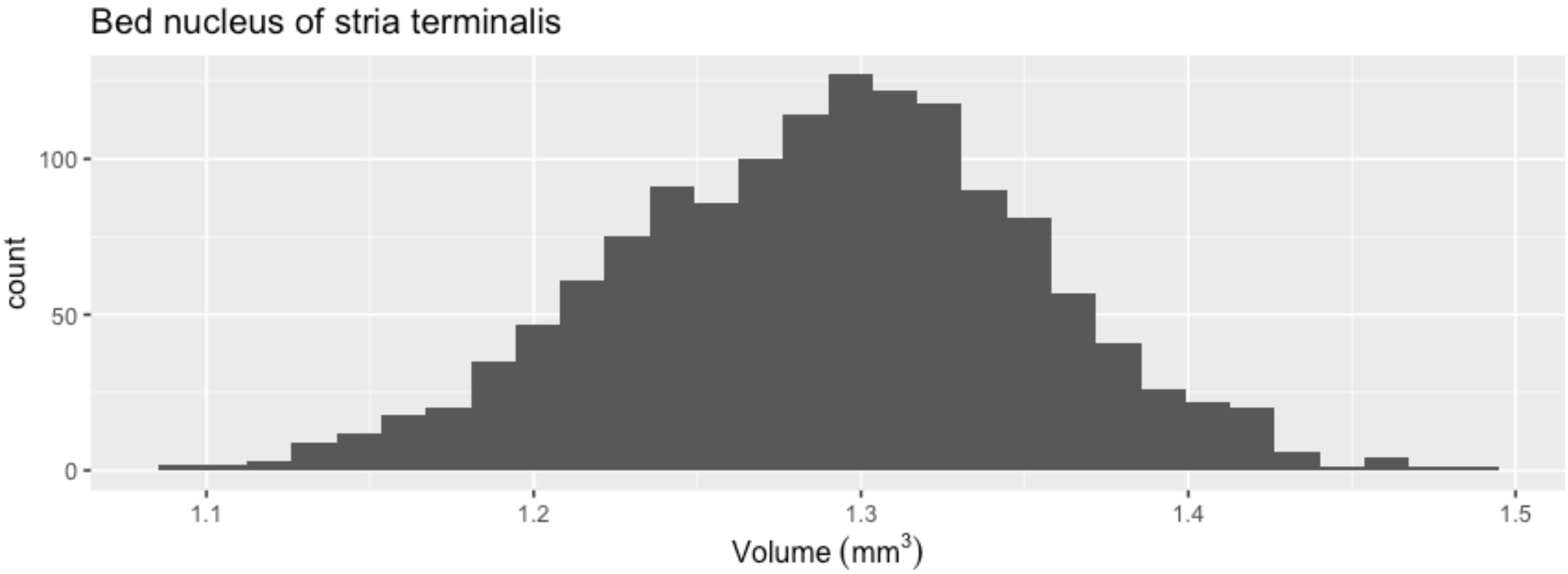
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Make it prettier

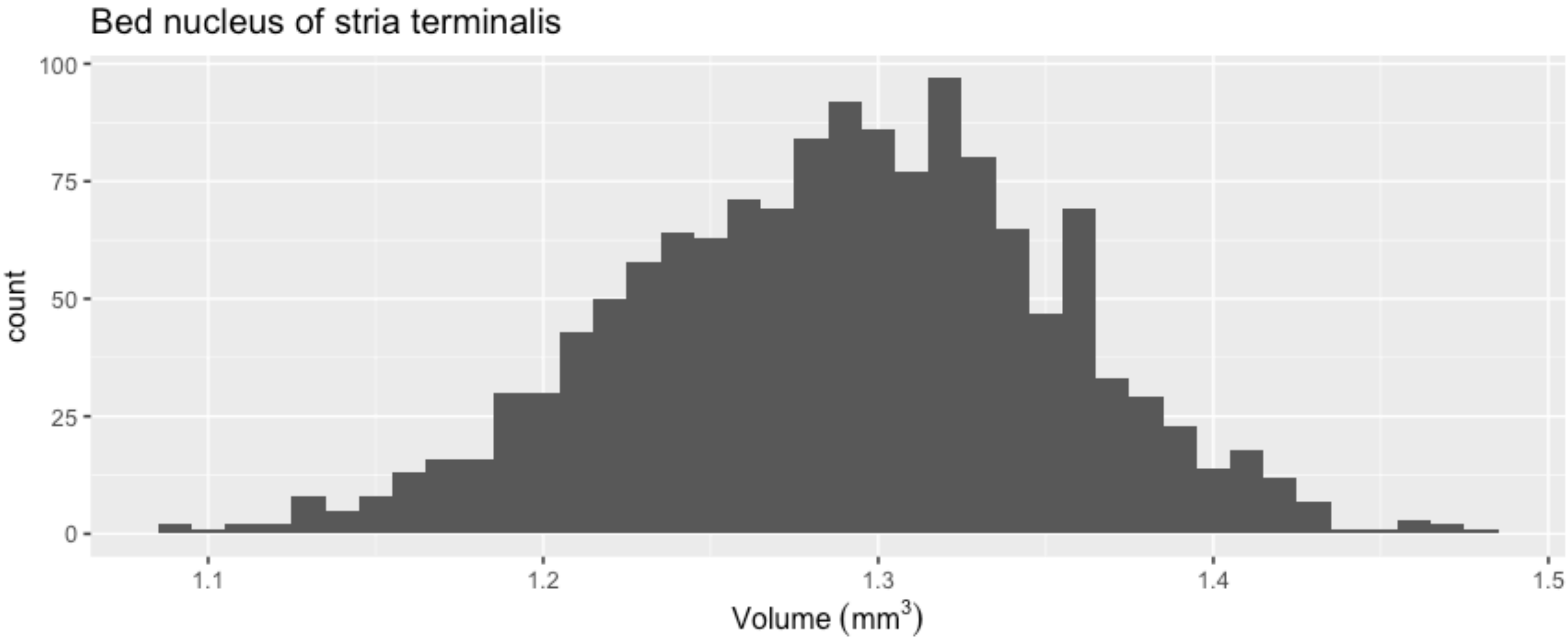
```
ggplot(mice) +
  aes(x=`bed nucleus of stria terminalis`) +
  geom_histogram() +
  xlab(bquote(Volume ~ (mm^3))) +
  ggtitle("Bed nucleus of stria terminalis") +
  theme_gray(16)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



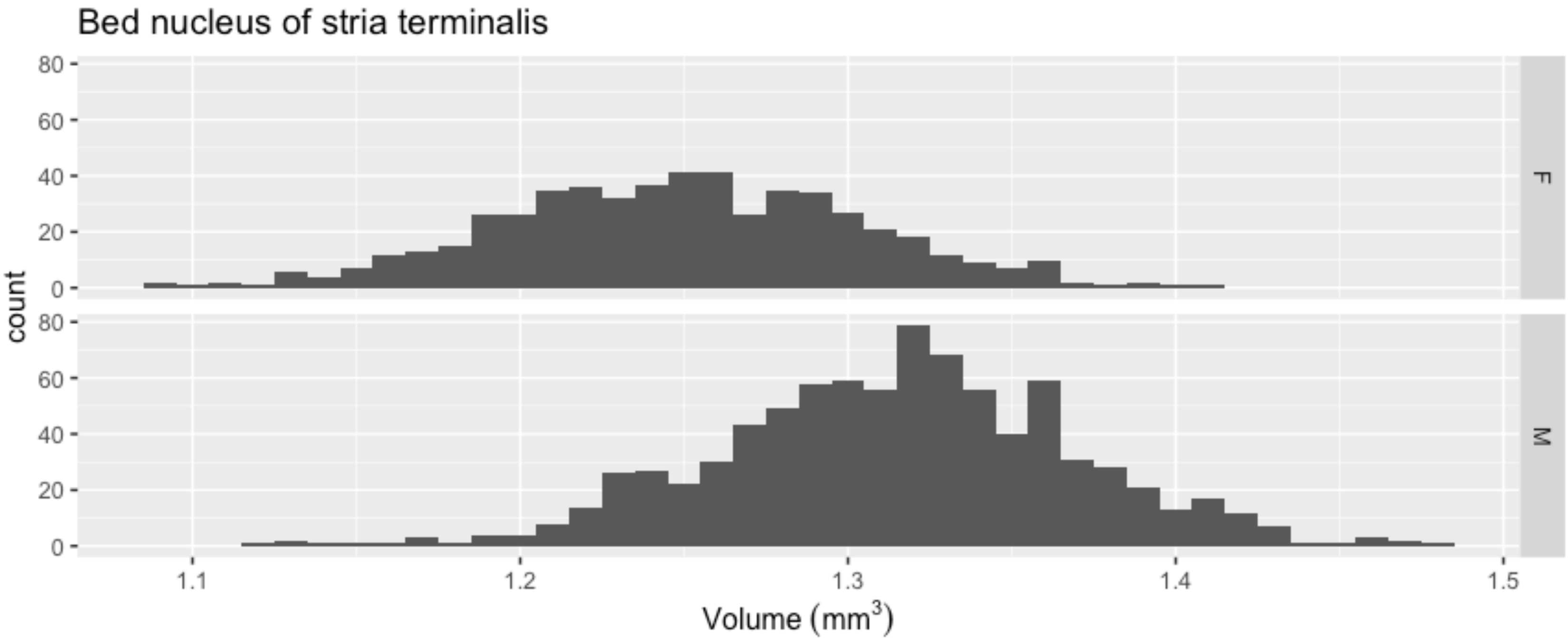
Histogram bins

```
ggplot(mice) +
  aes(x=`bed nucleus of stria terminalis`) +
  geom_histogram(binwidth = 0.01) +
  xlab(bquote(Volume ~ (mm^3))) +
  ggtitle("Bed nucleus of stria terminalis") +
  theme_gray(16)
```



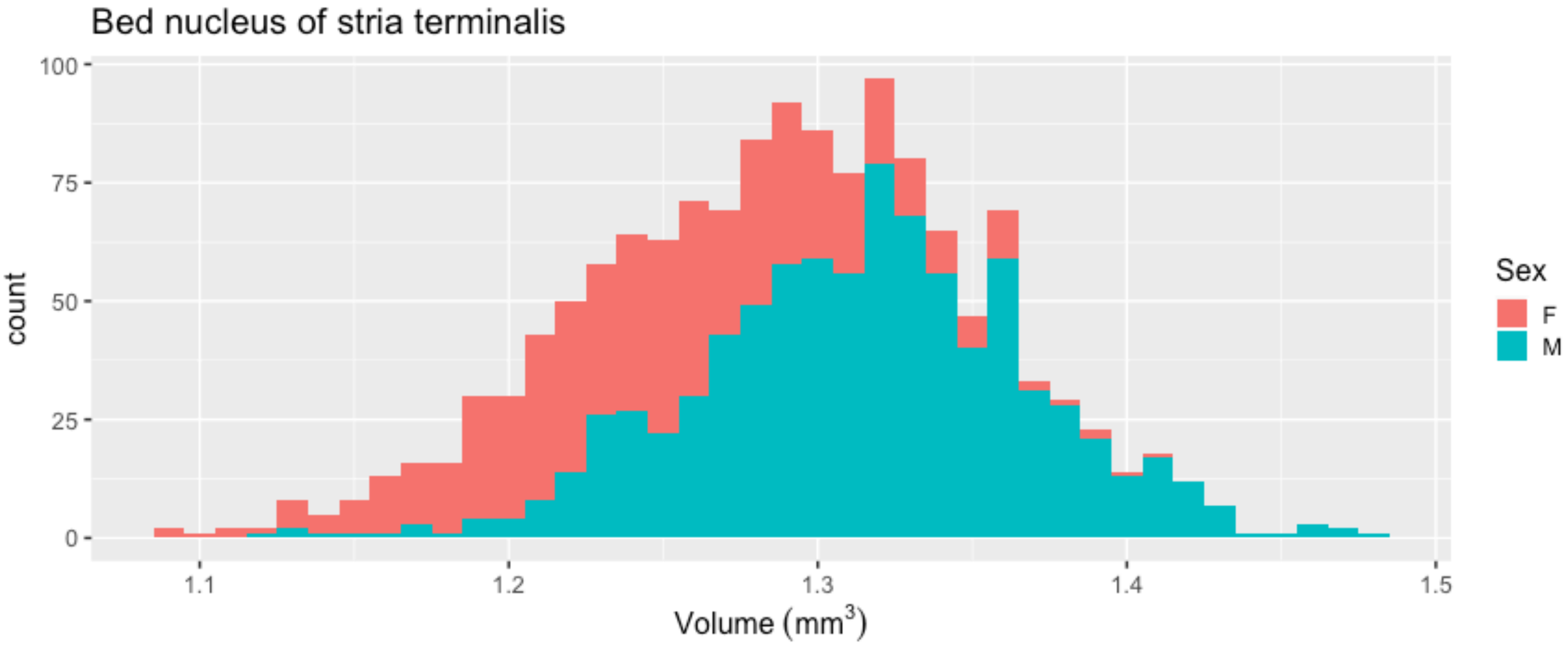
Facets

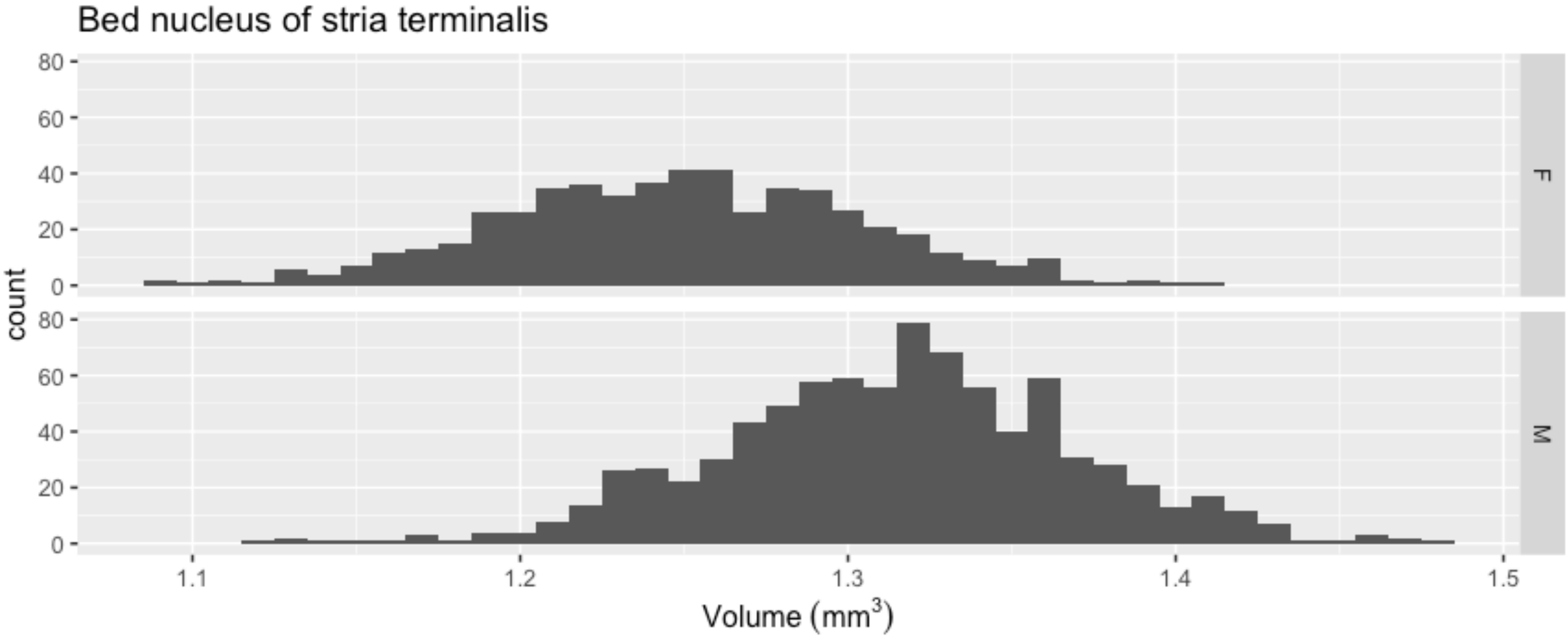
```r
ggplot(mice) +
  aes(x=`bed nucleus of stria terminalis`) +
  geom_histogram(binwidth = 0.01) +
  xlab(bquote(Volume ~ (mm^3))) +
  ggtitle("Bed nucleus of stria terminalis") +
  theme_gray(16) +
  facet_grid(Sex ~ .)
```



Colours

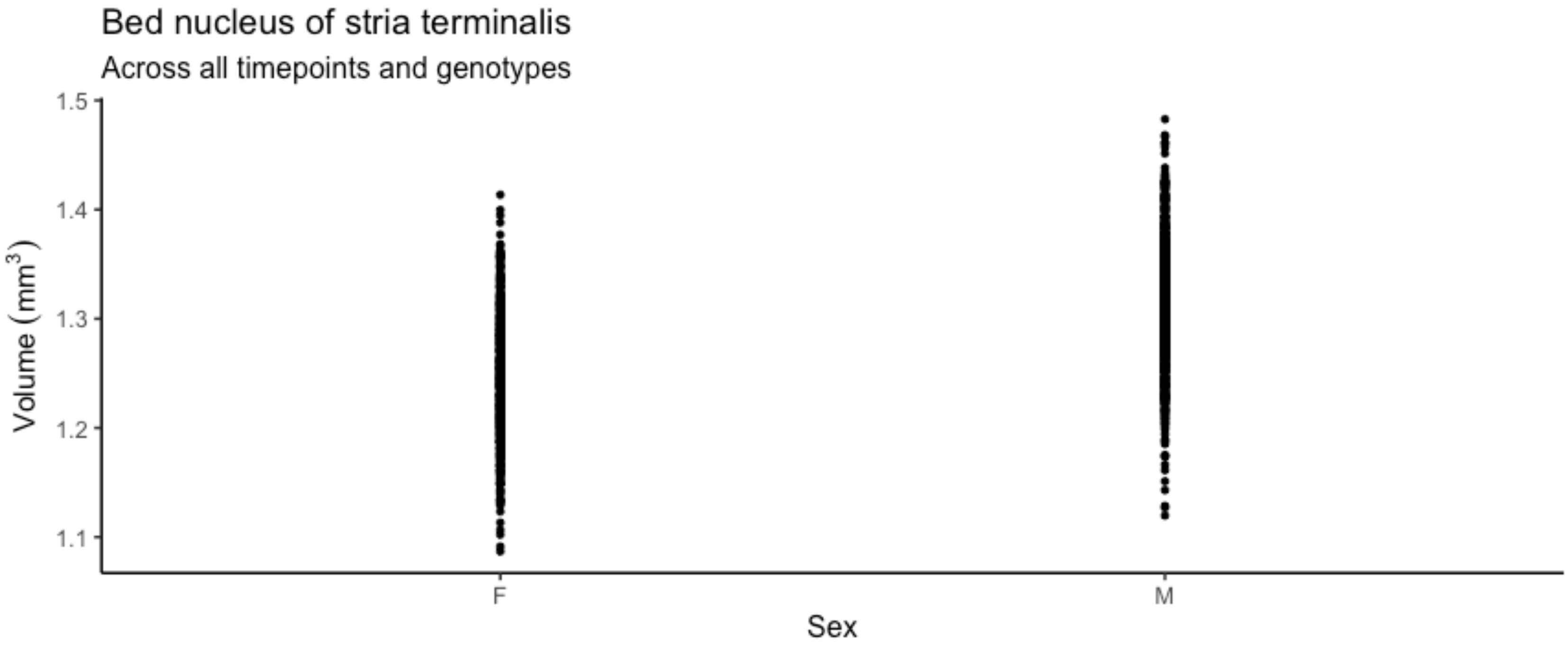
```r
ggplot(mice) +
  aes(x=`bed nucleus of stria terminalis`, fill=Sex) +
  geom_histogram(binwidth = 0.01) +
  xlab(bquote(Volume ~ (mm^3))) +
  ggtitle("Bed nucleus of stria terminalis") +
  theme_gray(16)
```




Bed nucleus of stria terminalis
count
80
60
40
20
0
F
M
1.1
1.2
1.3
1.4
1.5
Volume (mm³)

Points

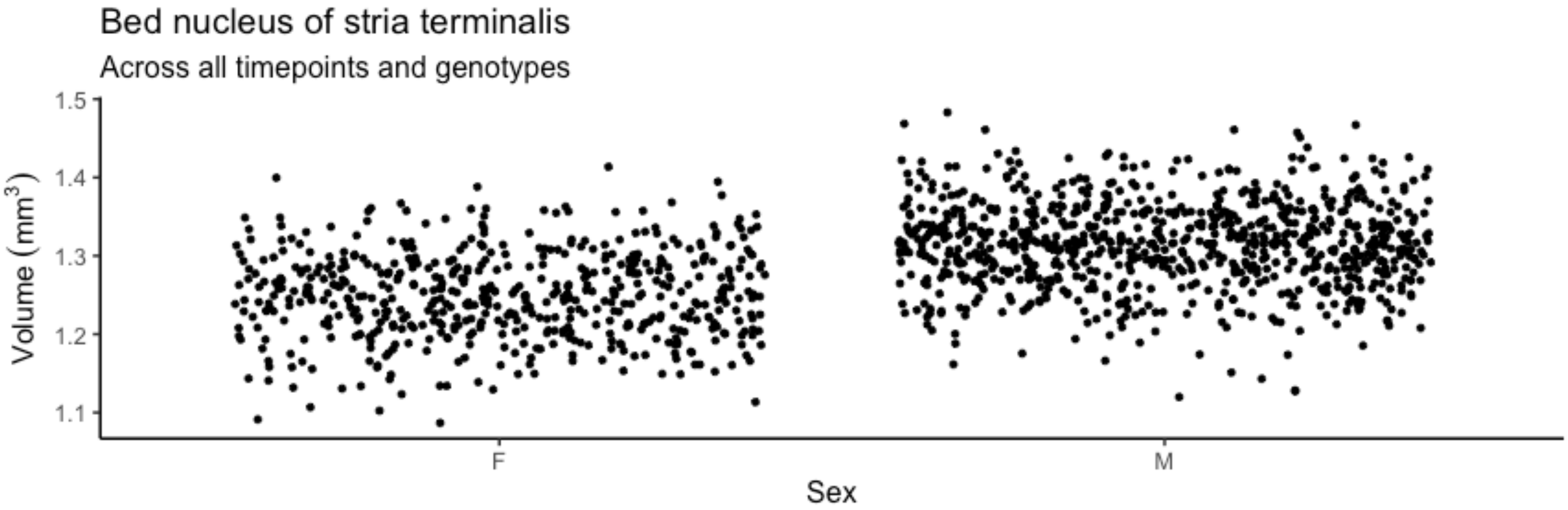
```
ggplot(mice) +
  aes(x=Sex, y=`bed nucleus of stria terminalis`) +
  geom_point() +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  theme_classic(16)
```



Points

That's not very useful - too many points to see separation.

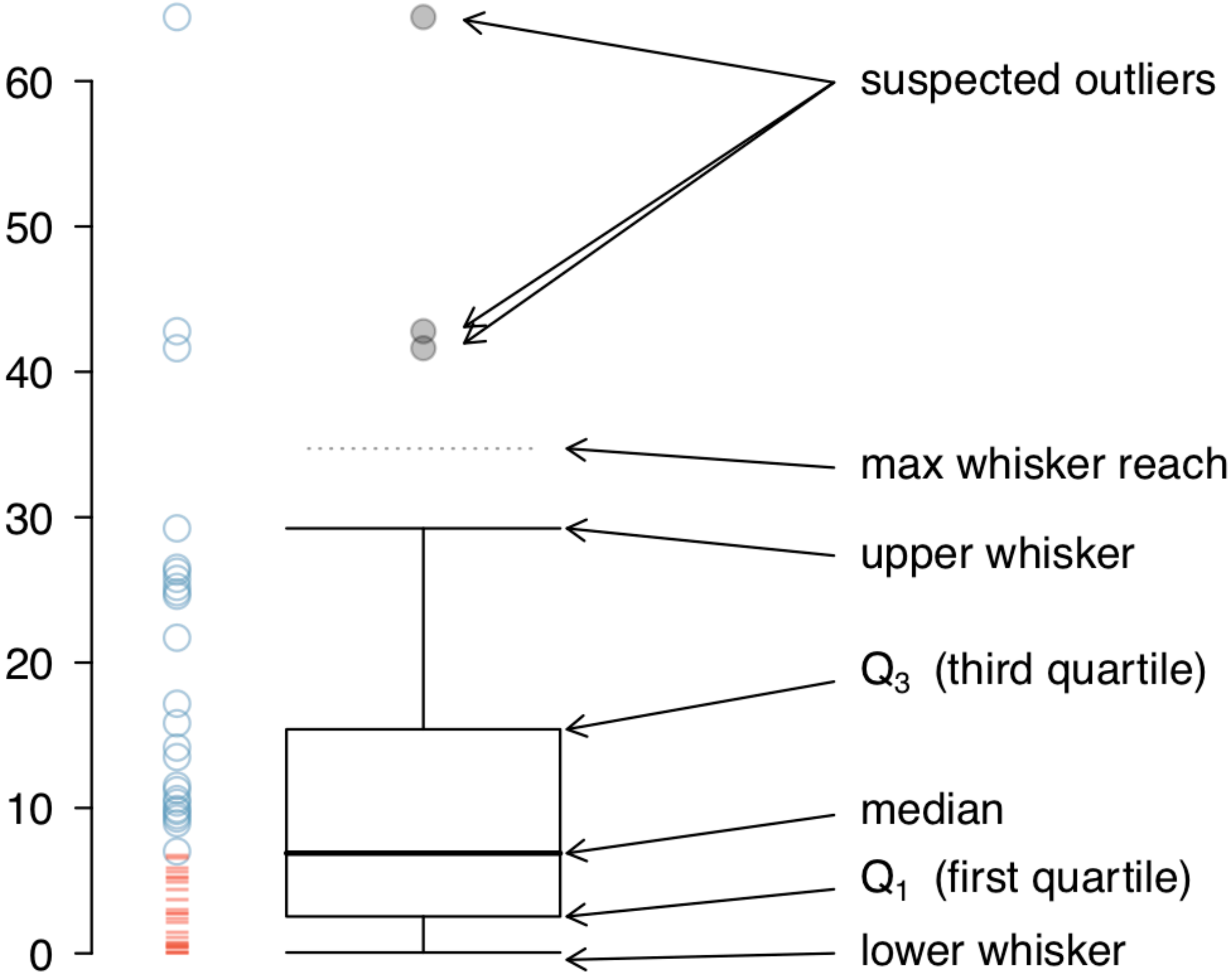
```
ggplot(mice) +
  aes(x=Sex, y=`bed nucleus of stria terminalis`) +
  geom_jitter() +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  theme_classic(16)
```



Boxplot

Good view of data distribution

```
ggplot(mice) +
  aes(x=Sex, y=`bed nucleus of stria terminalis`) +
  geom_boxplot() +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  theme_classic(16)
```


suspected outliers
max whisker reach
upper whisker
Q_3 (third quartile)
median
Q_1 (first quartile)
lower whisker
60
50
40
30
20
10
0

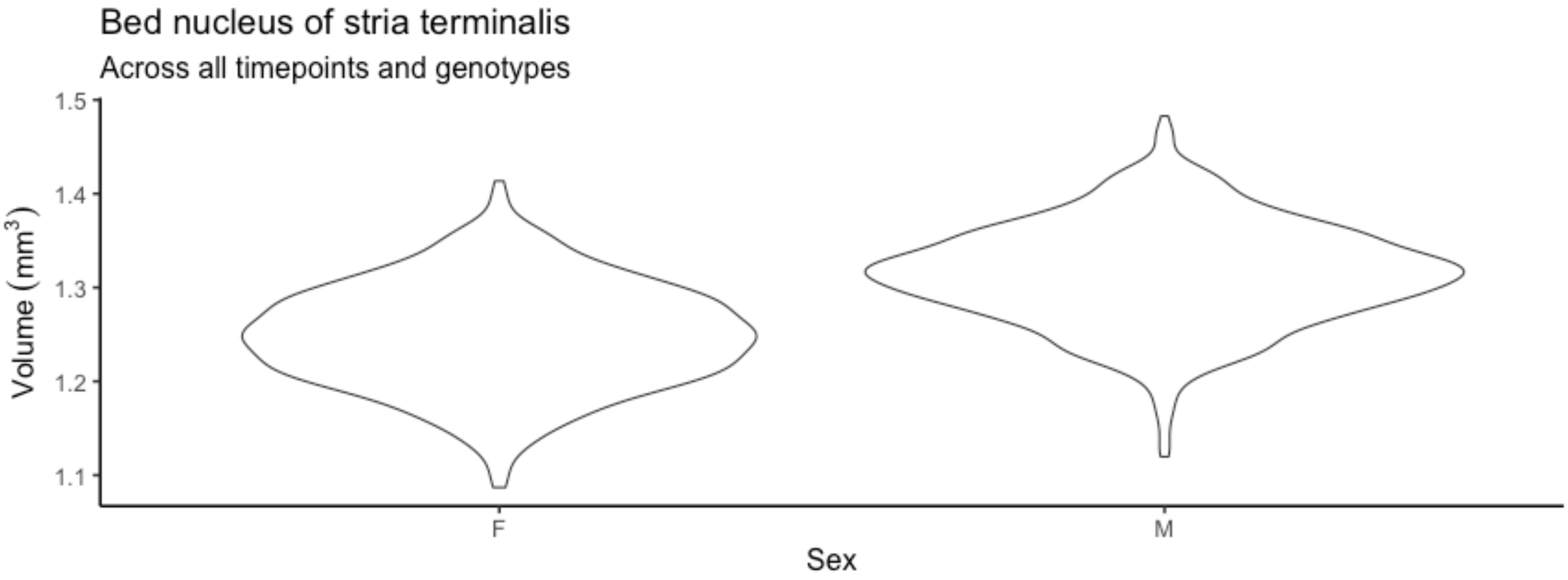
Ridge lines

```r
suppressMessages(library(ggridges))
ggplot(mice) +
  aes(y=Sex, x=`bed nucleus of stria terminalis`) +
  geom_density_ridges() +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  xlab(bquote(Volume ~ (mm^3))) +
  theme_classic(16)
```

```
## Picking joint bandwidth of 0.0132
```

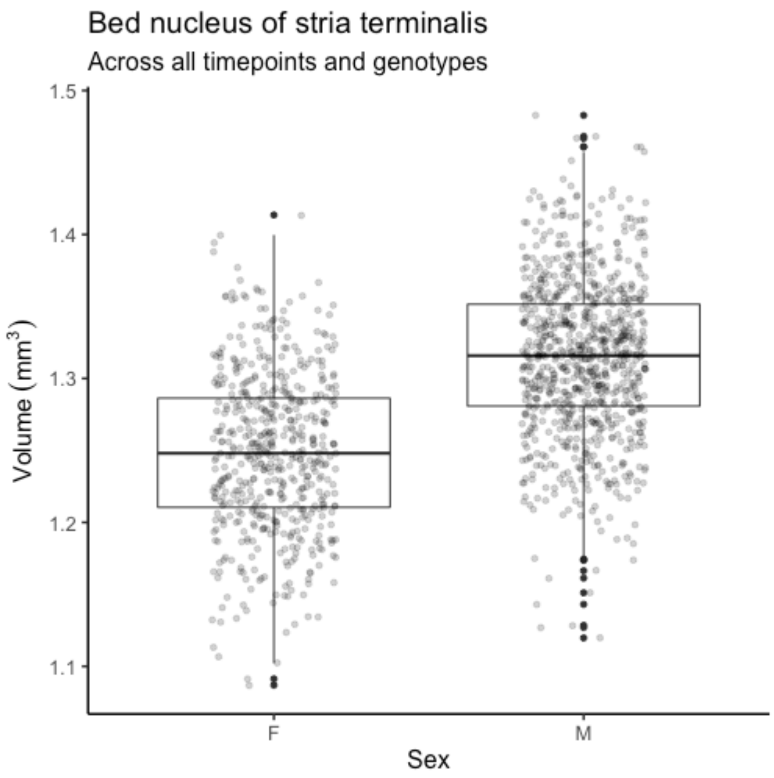
Violins

```
ggplot(mice) +
  aes(x=Sex, y=`bed nucleus of stria terminalis`) +
  geom_violin() +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  theme_classic(16)
```



Combining plot types

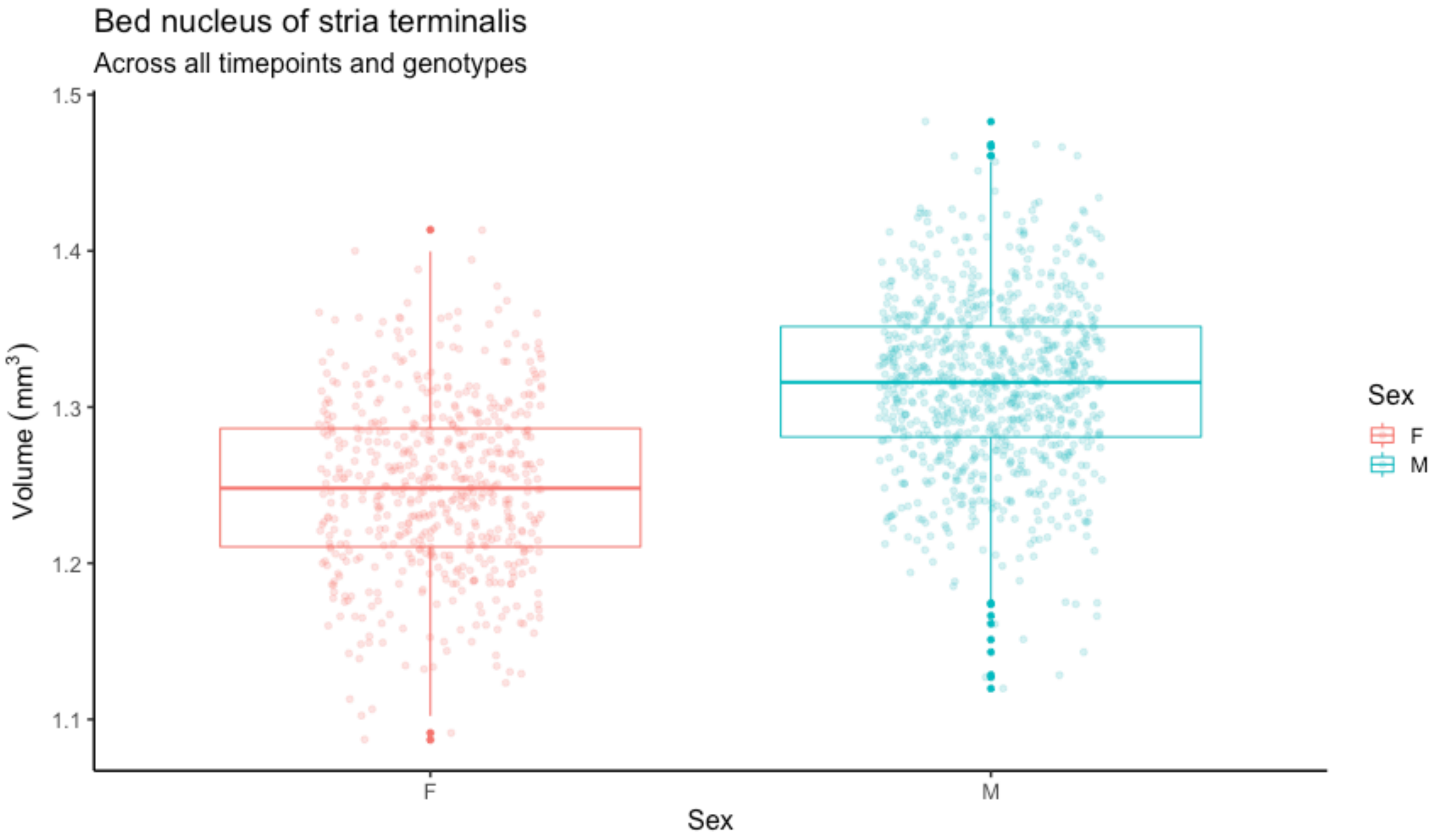
```r
ggplot(mice) +
  aes(x=Sex,
      y=`bed nucleus of stria terminalis`
      ) +
  geom_boxplot() +
  geom_jitter(width=0.2,
              alpha=0.2) +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  ylab(bquote(
    Volume ~ (mm^3))) +
  theme_classic(16)
```



Adding colour

```
ggplot(mice) +
  aes(x=Sex,
      y=`bed nucleus of stria terminalis`,
      colour=Sex) +
  geom_boxplot() +
  geom_jitter(width=0.2,
              alpha=0.2) +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all timepoints and genotypes") +
  ylab(bquote(
    Volume ~ (mm^3))) +
  theme_classic(16)
```

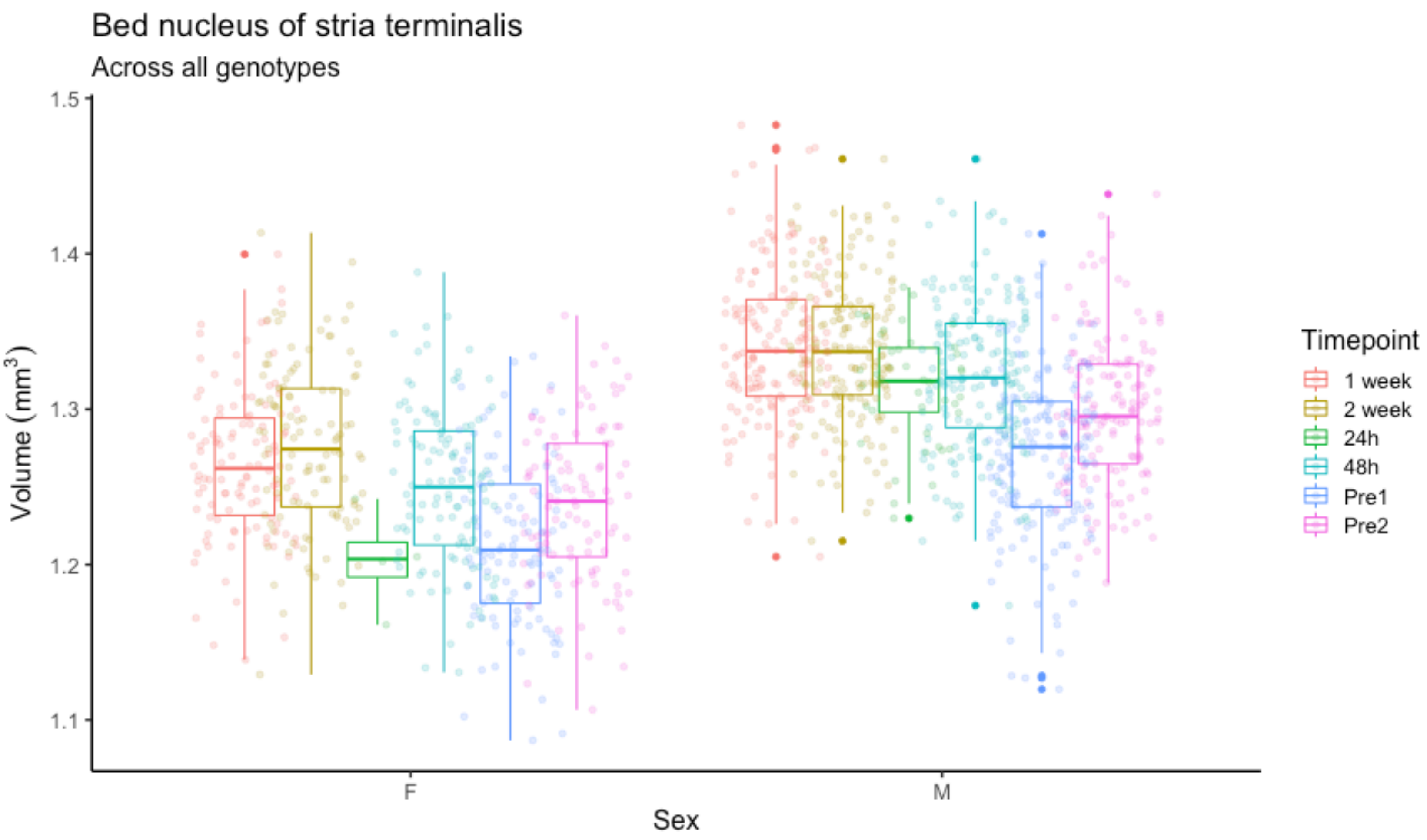
Adding colour



Using colour for additional information

```
ggplot(mice) +
  aes(x=Sex,
      y=`bed nucleus of stria terminalis`,
      colour=Timepoint) +
  geom_boxplot() +
  geom_jitter(alpha=0.2,
              position = position_jitterdodge(jitter.width = 0.2)) +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  theme_classic(16)
```

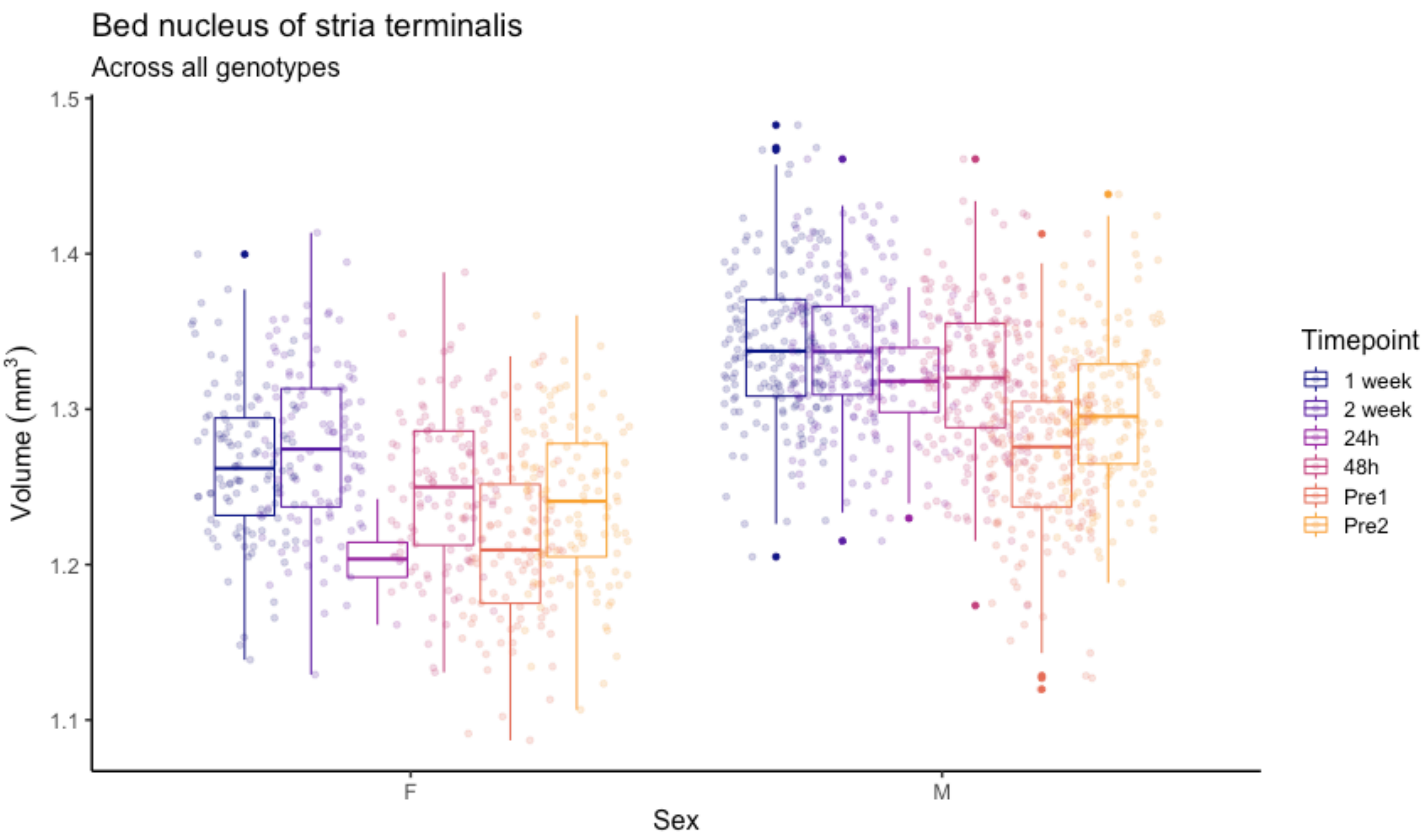
Using colour for additional information



Using colour for additional information

```
ggplot(mice) +
  aes(x=Sex,
      y=`bed nucleus of stria terminalis`,
      colour=Timepoint) +
  geom_boxplot() +
  geom_jitter(alpha=0.2,
              position = position_jitterdodge(jitter.width = 0.2)) +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  scale_colour_viridis_d(option="C", end=0.8) +
  theme_classic(16)
```

Using colour for additional information



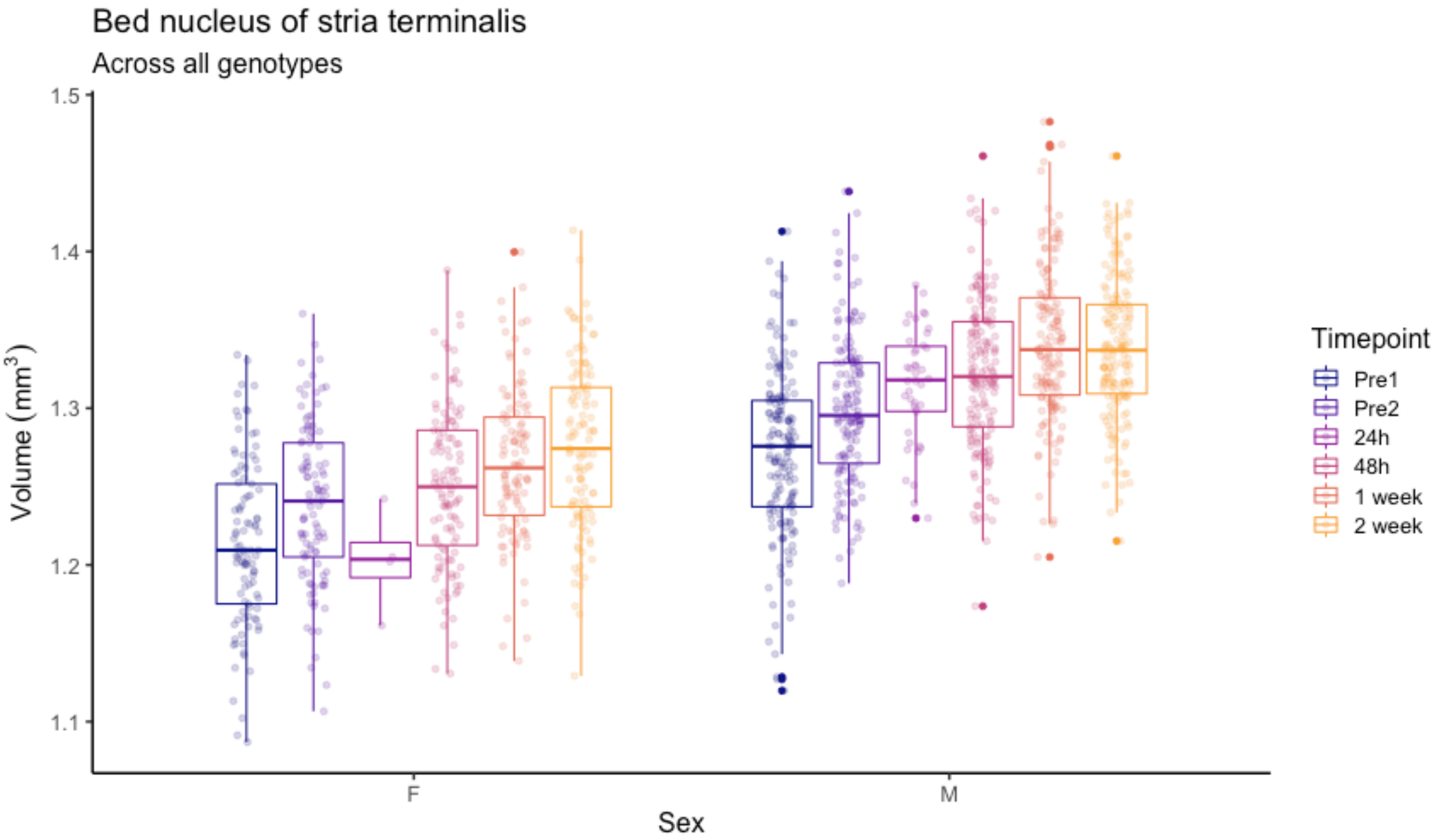
Factor order, again

Apparently the factor ordering was lost in data joining?

```r
mice <- mice %>%
  mutate(Timepoint=fct_relevel(Timepoint, "Pre1", "Pre2", "24h",
                               "48h", "1 week", "2 week"))

ggplot(mice) +
  aes(x=Sex,
      y=`bed nucleus of stria terminalis`,
      colour=Timepoint) +
  geom_boxplot() +
  geom_jitter(alpha=0.2,
              position = position_jitterdodge(jitter.width = 0.2)) +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle="Across all genotypes") +
  ylab(bquote(Volume ~ (mm^3))) +
  scale_colour_viridis_d(option="C", end=0.8) +
  theme_classic(16)
```

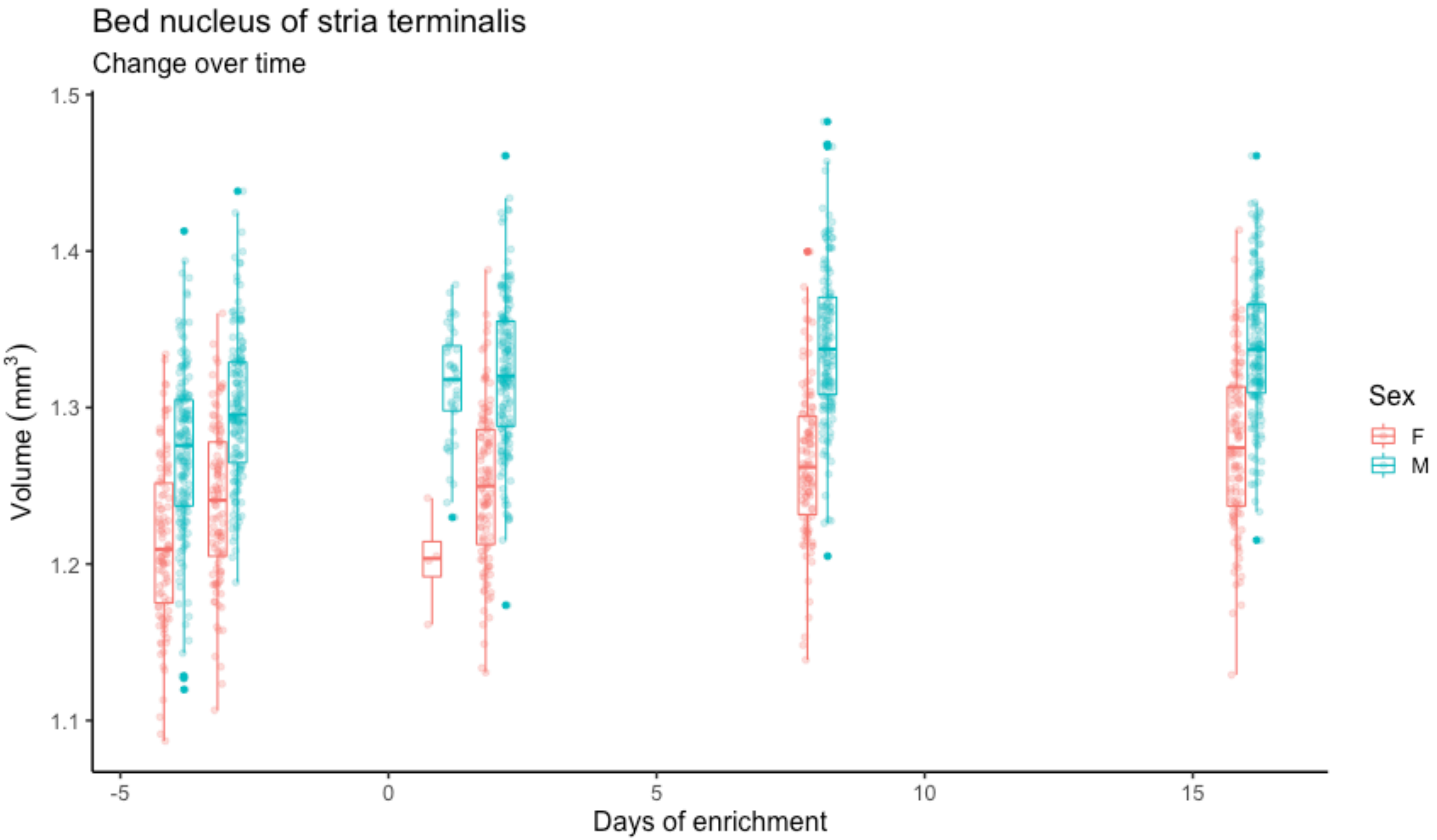
Factor ordering, again



Better encoding of time

```
ggplot(mice) +
  aes(x=DaysOfEE,
      y=`bed nucleus of stria terminalis`,
      colour=Sex) +
  geom_boxplot(aes(group=interaction(Timepoint, Sex))) +
  geom_jitter(alpha=0.25, position =
                position_jitterdodge(jitter.width = 0.2)) +
  ylab(bquote(Volume ~ (mm^3))) +
  xlab("Days of enrichment") +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle = "Change over time") +
  theme_classic(16)
```

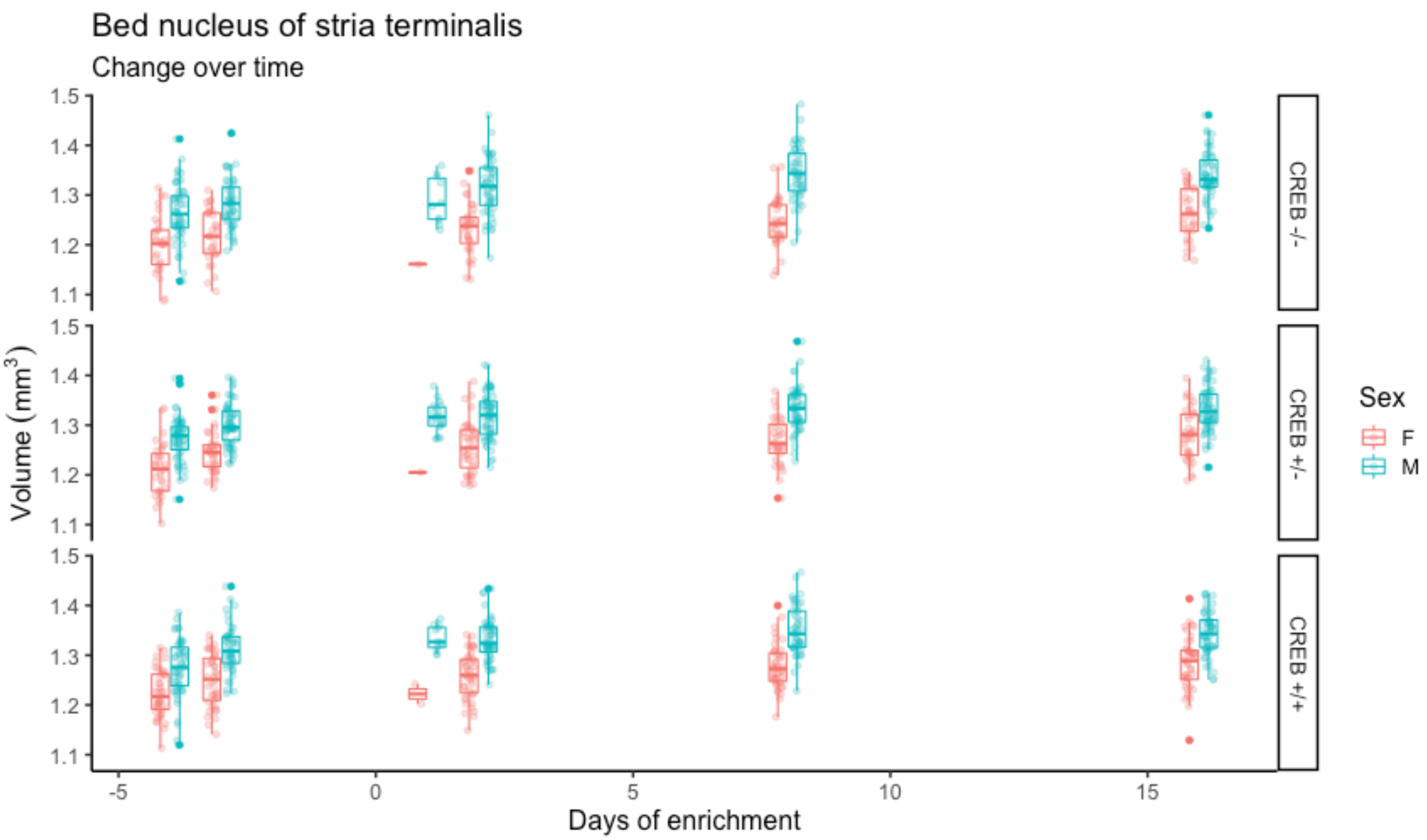
Better encoding of time



Combining colours and facets

```r
ggplot(mice) +
  aes(x=DaysOfEE,
      y=`bed nucleus of stria terminalis`,
      colour=Sex) +
  geom_boxplot(aes(group=interaction(Timepoint, Sex))) +
  geom_jitter(alpha=0.25, position =
                position_jitterdodge(jitter.width = 0.2)) +
  ylab(bquote(Volume ~ (mm^3))) +
  xlab("Days of enrichment") +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle = "Change over time") +
  facet_grid(Genotype ~ .) +
  theme_classic(16)
```

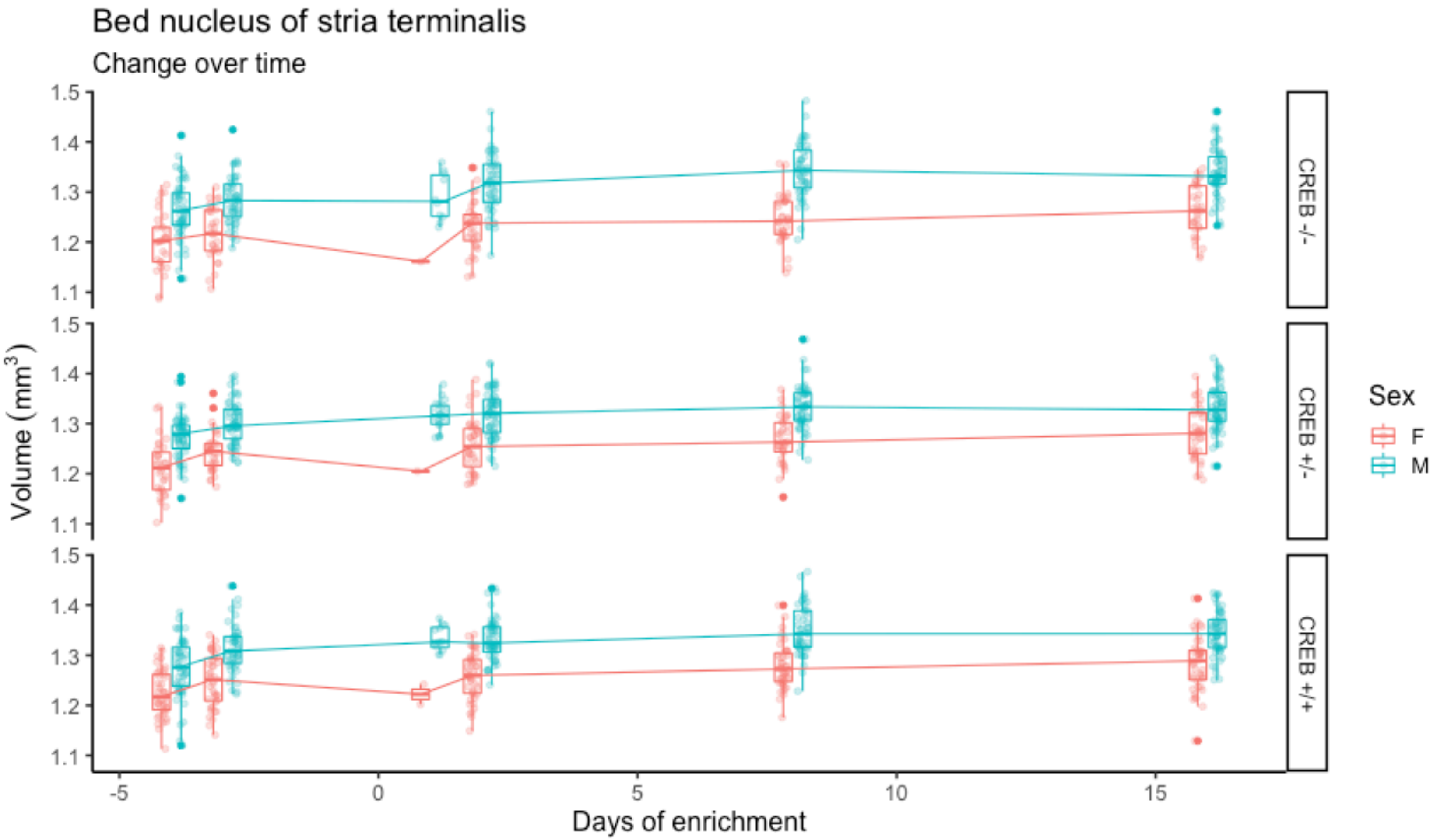
Combining colours and facets



Adding lines

```r
ggplot(mice) +
  aes(x=DaysOfEE,
      y=`bed nucleus of stria terminalis`,
      colour=Sex) +
  geom_boxplot(aes(group=interaction(Timepoint, Sex))) +
  geom_jitter(alpha=0.25, position =
                position_jitterdodge(jitter.width = 0.2)) +
  stat_summary(fun.y = median, geom="line",
               position =
                 position_jitterdodge(jitter.width = 0.2)) +
  ylab(bquote(Volume ~ (mm^3))) +
  xlab("Days of enrichment") +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle = "Change over time") +
  facet_grid(Genotype ~ .) +
  theme_classic(16)
```

Adding lines



Descriptive statistics

Descriptive statistics

summarize datapoints into measures of

- central tendency
  - mean/average: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - median: sort numbers, pick middle
  - mode: most common number
- variance/dispersion
  - standard deviation: $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$
  - range: smallest to largest value
  - interquartile range: recursive median calculations (median of upper half, median of lower half)

Plotting mean and standard deviation

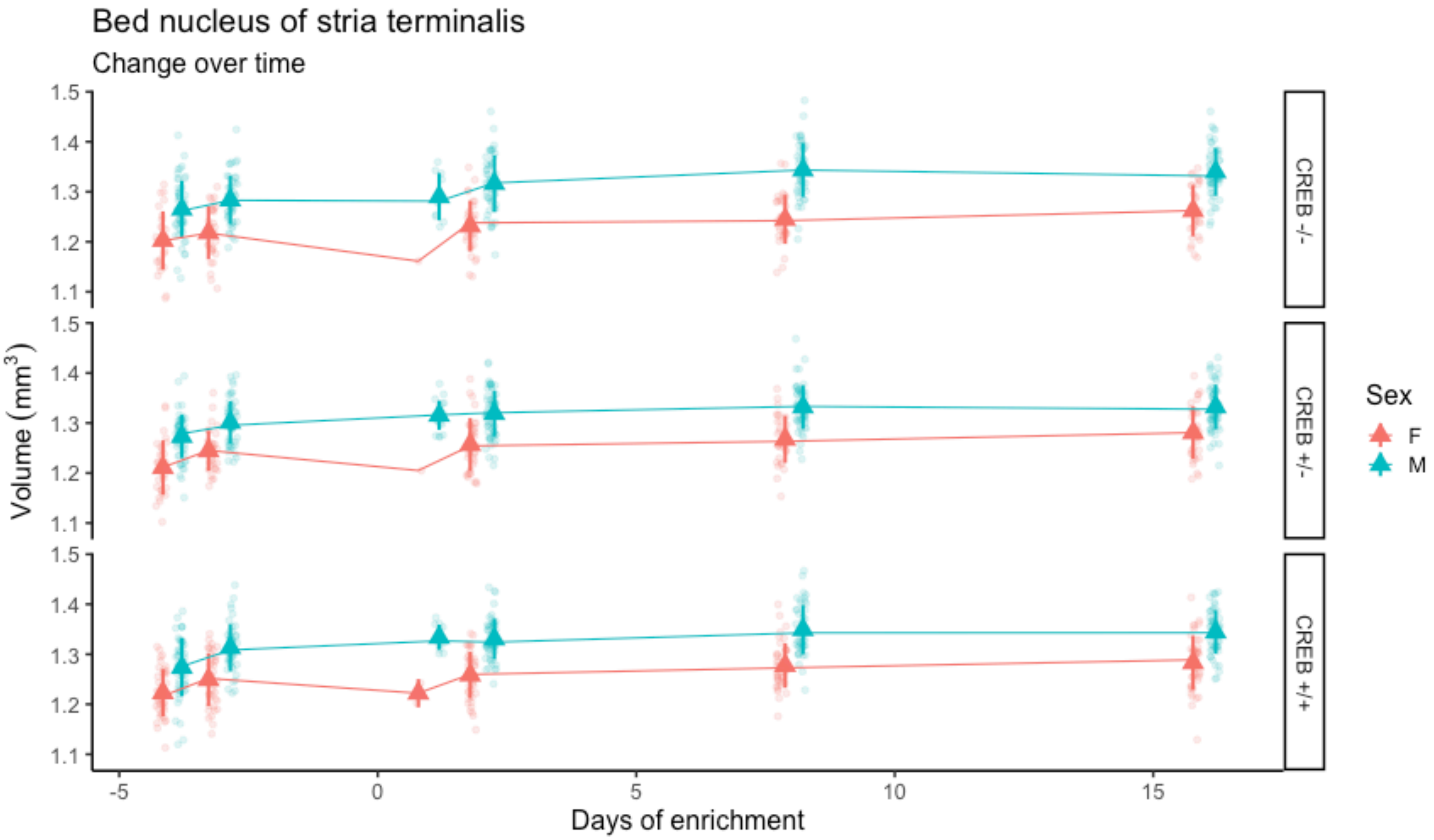
```r
suppressMessages(library(Hmisc))

p <- position_jitterdodge(jitter.width = 0.2)

ggplot(mice) +
  aes(x=DaysOfEE,
      y=`bed nucleus of stria terminalis`,
      colour=Sex) +
  stat_summary(fun.data=mean_sdl, geom="pointrange", size=1,
               shape=17, fun.args=(mult=1), position=p) +
  geom_jitter(alpha=0.15, position = p) +
  stat_summary(fun.y = median, geom="line",
               position =p) +
  ylab(bquote(Volume ~ (mm^3))) +
  xlab("Days of enrichment") +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle = "Change over time") +
  facet_grid(Genotype ~ .) +
  theme_classic(16)
```

```
## Warning: Removed 2 rows containing missing values (geom_pointrange).
```

Plotting mean and standard deviation



Standard error of the mean

Standard deviations independent of sample size. Statistical tests usually take sample size into account. The standard error of the mean is the standard deviation divided by the square root of n.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Plotting mean and standard error

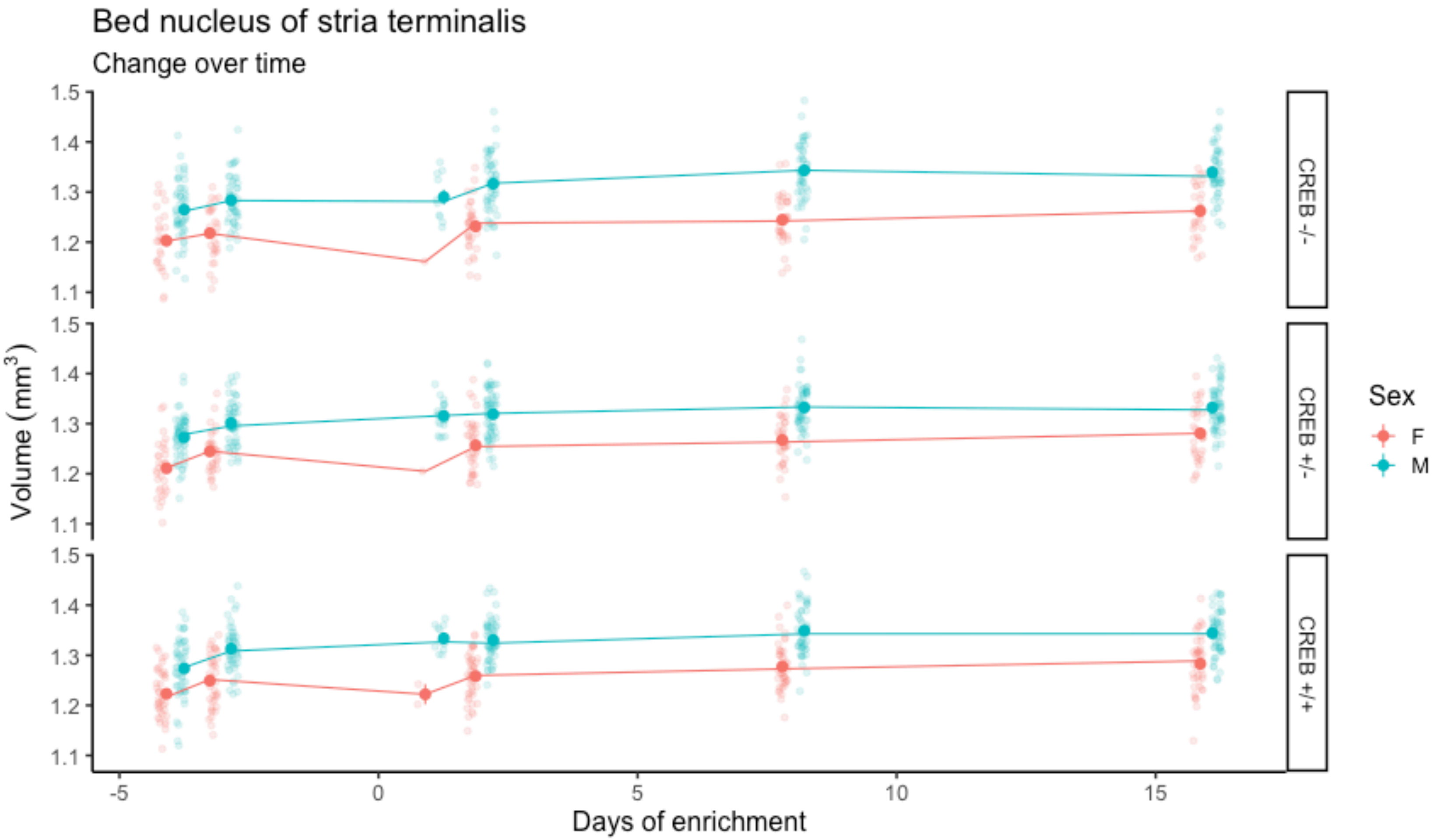
```r
suppressMessages(library(Hmisc))

p <- position_jitterdodge(jitter.width = 0.2)

ggplot(mice) +
  aes(x=DaysOfEE,
      y=`bed nucleus of stria terminalis`,
      colour=Sex) +
  stat_summary(fun.data=mean_se, geom="pointrange",
               fun.args=(mult=1), position=p) +
  geom_jitter(alpha=0.15, position = p) +
  stat_summary(fun.y = median, geom="line",
                position =p) +
  ylab(bquote(Volume ~ (mm^3))) +
  xlab("Days of enrichment") +
  ggtitle("Bed nucleus of stria terminalis",
          subtitle = "Change over time") +
  facet_grid(Genotype ~ .) +
  theme_classic(16)
```

```
## Warning: Removed 2 rows containing missing values (geom_pointrange).
```

Plotting mean and standard error



Summary table

```
mice %>%
  group_by(Genotype, Sex, DaysOfEE) %>%
  summarise(m=mean(`bed nucleus of stria terminalis`)) %>%
  spread(DaysOfEE, m)  %>%
  knitr::kable(format = 'html')
```

Genotype	Sex	-4	-3	1	2	8	16
CREB -/-	F	1.202775	1.217989	1.161297	1.231568	1.244746	1.262104
CREB -/-	M	1.264994	1.283486	1.290038	1.316406	1.343368	1.339794
CREB +/-	F	1.210892	1.244424	1.205037	1.256917	1.267488	1.280610
CREB +/-	M	1.272896	1.301316	1.315517	1.318961	1.332175	1.332277
CREB +/+	F	1.223120	1.249269	1.222168	1.258373	1.277564	1.283006
CREB +/+	M	1.273769	1.313224	1.333948	1.330533	1.349106	1.344397

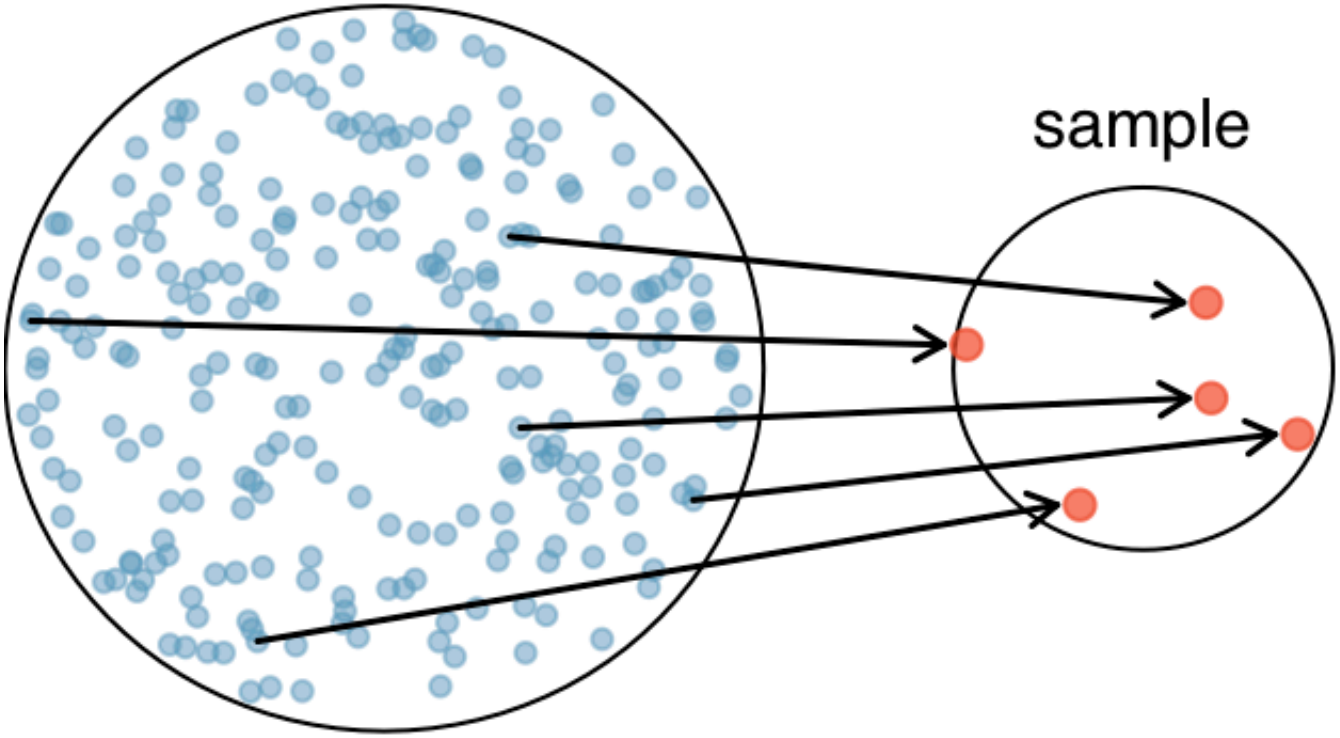
Summary table

```
mice %>%
  group_by(Genotype, Sex, DaysOfEE) %>%
  summarise(m=paste(round(mean(`bed nucleus of stria terminalis`),2),
                    round(sd(`bed nucleus of stria terminalis`), 2), sep
  spread(DaysOfEE, m) %>%
  knitr::kable(format = 'html')
```

Genotype	Sex	-4	-3	1	2	8	16
CREB -/-	F	1.2±0.06	1.22±0.05	1.16±NA	1.23±0.05	1.24±0.05	1.26±0.05
CREB -/-	M	1.26±0.06	1.28±0.05	1.29±0.05	1.32±0.06	1.34±0.05	1.34±0.05
CREB +/-	F	1.21±0.05	1.24±0.04	1.21±NA	1.26±0.05	1.27±0.05	1.28±0.05
CREB +/-	M	1.27±0.04	1.3±0.04	1.32±0.03	1.32±0.05	1.33±0.04	1.33±0.04
CREB +/+	F	1.22±0.05	1.25±0.05	1.22±0.03	1.26±0.05	1.28±0.04	1.28±0.05
CREB +/+	M	1.27±0.06	1.31±0.05	1.33±0.02	1.33±0.04	1.35±0.05	1.34±0.04

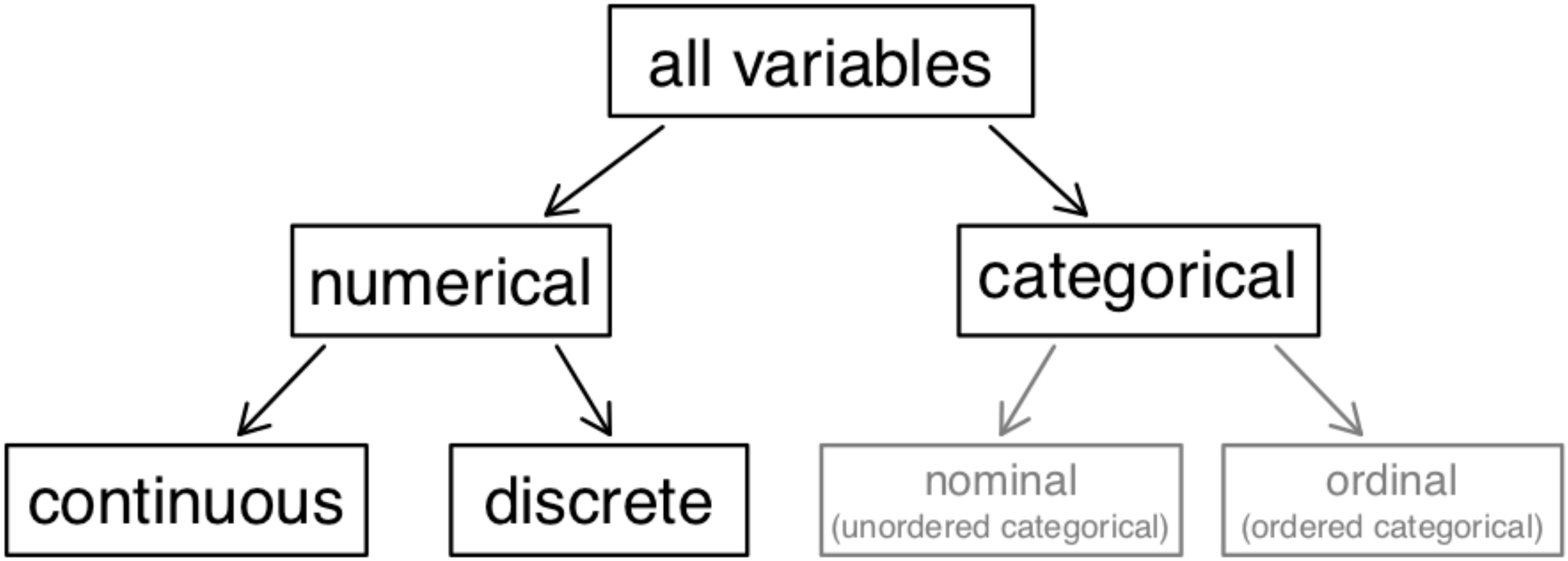
Statistical tests

From populations to samples



OpenIntro Statistics, Diez, Barr, and Cetinkaya-Rundel, 2015

Data types



Data types determine choice of statistics and/or encoding.

OpenIntro Statistics, Diez, Barr, and Cetinkaya-Rundel, 2015

Sex ratios

Are the sex ratios in our data balanced?

```
baseline <- mice %>% filter(Timepoint == "Pre1")
addmargins(with(baseline, table(Sex)))
```

```
## Sex
##   F   M Sum
## 101 165 266
```

What should we expect?

Assume equal probability of male or female

```
nrow(baseline) / 2
```

```
## [1] 133
```

How likely was our real value?

Binomial distribution - flip of a coin.

```r
rbinom(1, 1, 0.5)
```

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

```r
rbinom(1, 1, 0.5)
```

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

```r
rbinom(1, 1, 0.5)
```

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

```r
rbinom(10, 1, 0.5)
```

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

How likely was our real value?

```
baseline <- mice %>% filter(Timepoint == "Pre1")
addmargins(with(baseline, table(Sex)))
```

```
## Sex
##   F   M Sum
## 101 165 266
```

Assuming random choice of male or female:

```
distribution <- rbinom(266, 1, 0.5)
sum(distribution==1)
```

```
## [1] 139
```

```
rbinom(1, 266, 0.5)
```

```
## [1] 121
```

Long run probability

We did a single experiment, and obtained 101 Females and 165 Males.

If we were to rerun the experiment again and again and again, and each experimental mouse had a 50/50 chance of being male or female, how often would we obtain 101 Females or fewer?

```r
nexperiments <- 1000
females <- vector(length=nexperiments)
for (i in 1:nexperiments) {
  females[i] <- rbinom(1, 266, 0.5)
}
head(females)
```

```
## [1] 132 138 135 136 125 140
```

Can be shortened as

```r
females2 <- rbinom(nexperiments, 266, 0.5)
head(females2)
```

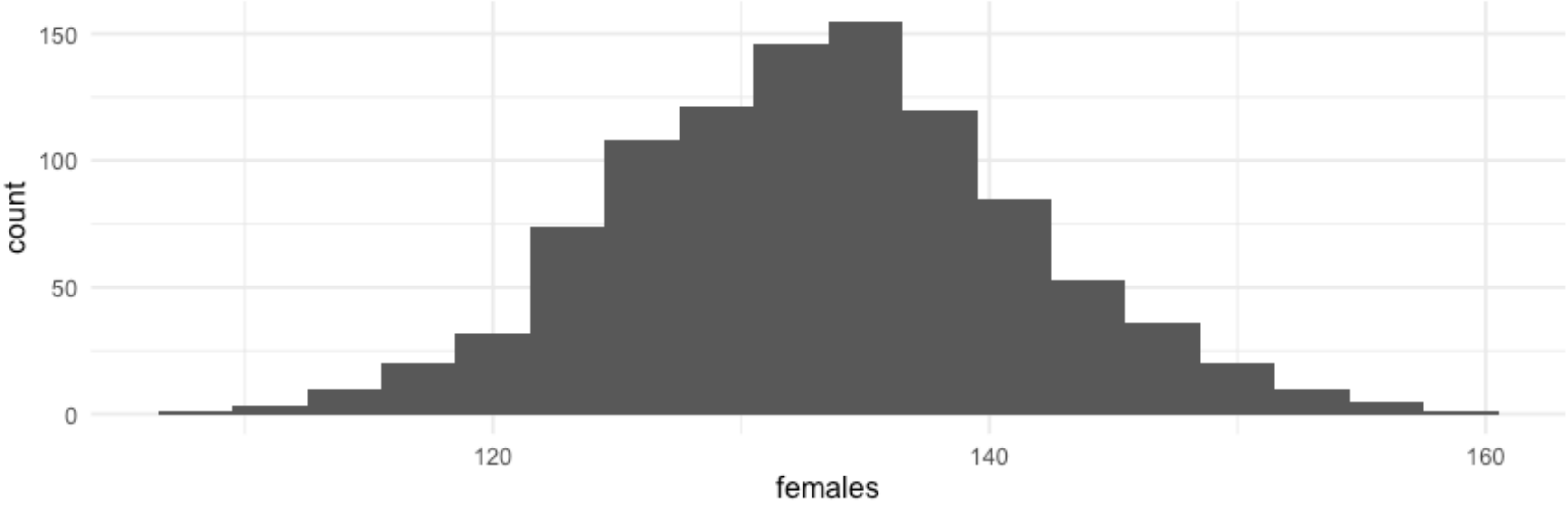
```
## [1] 125 129 149 139 131 133
```

Long run probability

```
head(females)
```

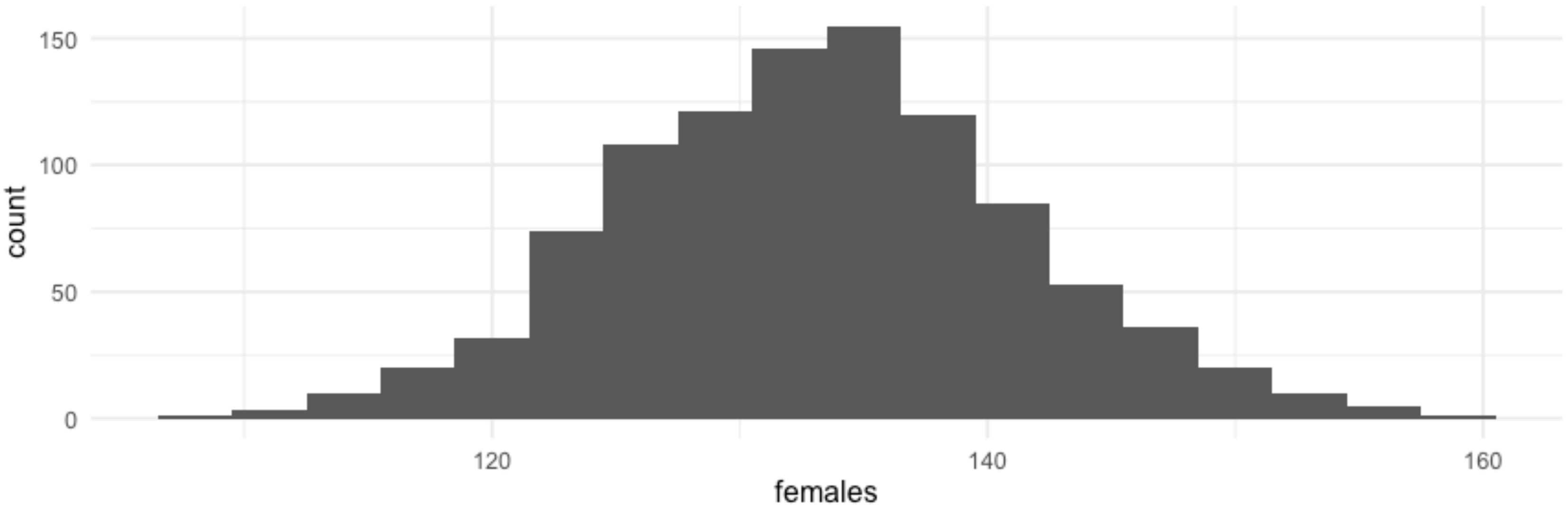
```
## [1] 132 138 135 136 125 140
```

```
ggplot(data.frame(females=females)) +
  aes(x=females) +
  geom_histogram(binwidth = 3) +
  theme_minimal(16)
```



Long run probability

```
ggplot(data.frame(females=females)) +
  aes(x=females) +
  geom_histogram(binwidth = 3) +
  theme_minimal(16)
```

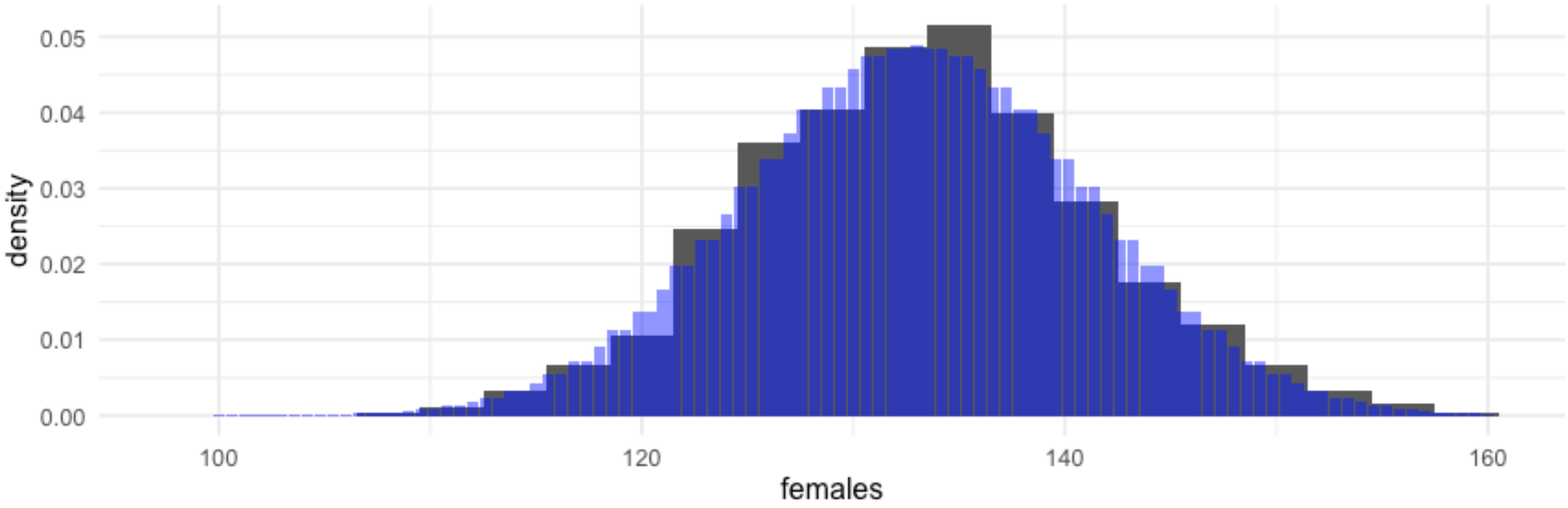


```
sum(females<=101)
```

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

Closed form solution

```
ggplot() +
  geom_histogram(data=data.frame(females=females),
                 aes(x=females, y=..density..),
                 binwidth = 3) +
  geom_bar(aes(c(100:160)), stat="function",
           fun=function(x) dbinom(round(x), 266, 0.5),
           alpha=0.5, fill="blue") +
  theme_minimal(16)
```



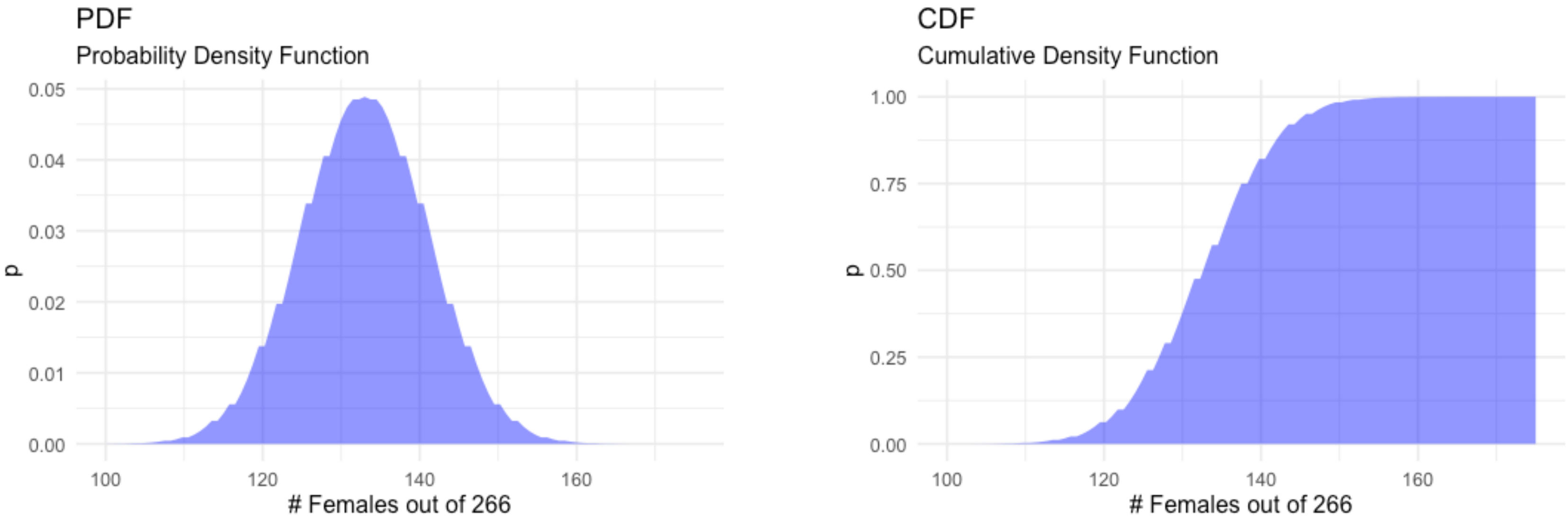
Closed form solution

```
pbinom(101, 266, 0.5)
```

```
## [1] 5.223361e-05
```

```
sum(dbinom(0:101, 266, 0.5))
```

```
## [1] 5.223361e-05
```



Review

- We asked whether the sex ratio in the study was likely to be random, assuming an equal chance of an experimental mouse being male or female.
- We simulated 1000 studies under the assumption of n=266 and the odds of being female = 50%
- This is the null hypothesis.
- Our random data simulations test the null hypothesis: what would happen if we ran the experiment again and again and again under the same conditions assuming random assignment of males and females?
- Our p-value - the long run probability under repeated experiments - was vanishingly small.

So the choice of sex was almost certainly non-random. Does it matter?

Contingency table

```
baseline <- mice %>% filter(Timepoint == "Pre1")
with(baseline, table(Sex, Genotype))
```

```
##    Genotype
## Sex CREB -/- CREB +/- CREB +/+
##   F       29       31       41
##   M       53       59       53
```

```
addmargins(with(baseline, table(Sex, Genotype)))
```

```
##      Genotype
## Sex   CREB -/- CREB +/- CREB +/+ Sum
##   F         29       31       41 101
##   M         53       59       53 165
##   Sum       82       90       94 266
```

What would we expect?

The table of observed numbers

```
addmargins(with(baseline, table(Sex, Genotype))) %>%
  knitr::kable(format = 'html')
```

	CREB -/-	CREB +/-	CREB +/+	Sum
F	29	31	41	101
M	53	59	53	165
Sum	82	90	94	266

Calculating the expected numbers

	CREB -/-	CRE +/-	CREB +/+	Sum
F	82*101/266	90*101/266	94*101/266	101
M	82*165/266	90*165/266	94*165/266	165
Sum	82	90	94	266

Using the chisq.test function for these calculations

```
xtest <- with(baseline, chisq.test(Sex, Genotype))
addmargins(xtest$observed)
```

```
##      Genotype
## Sex   CREB -/- CREB +/- CREB +/+ Sum
##   F         29       31       41 101
##   M         53       59       53 165
##   Sum       82       90       94 266
```

```
addmargins(xtest$expected)
```

```
##      Genotype
## Sex   CREB -/- CREB +/- CREB +/+ Sum
##   F   31.13534 34.17293 35.69173 101
##   M   50.86466 55.82707 58.30827 165
##   Sum 82.00000 90.00000 94.00000 266
```

χ^2 test

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij} - \tilde{n}_{ij}}{\tilde{n}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - \frac{n_i + n + j}{n})^2}{\frac{n_i + n + j}{n}}$$

		Y					
		y_1		y_j		y_l	Total (rows)
	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	x_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	Total (columns)	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

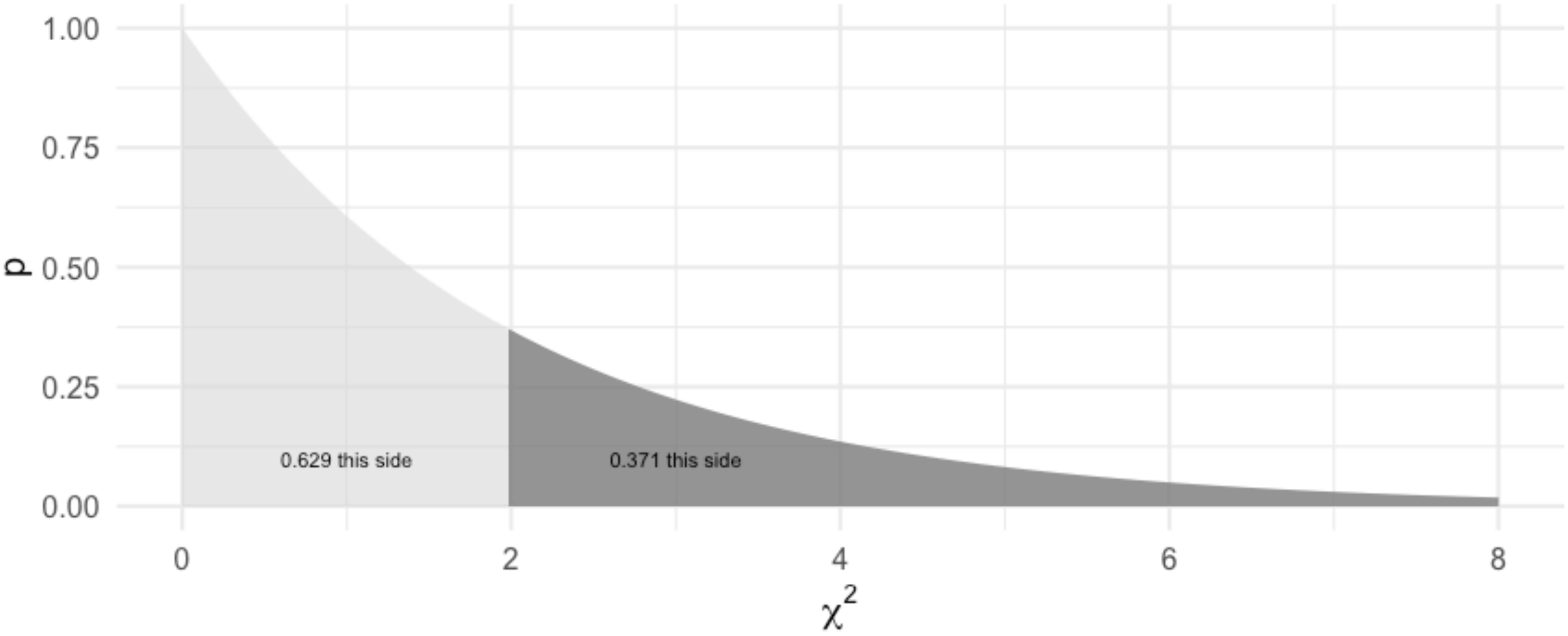
```
sum( ((xtest$observed - xtest$expected)^2)/xtest$expected )
```

χ^2 test

```
sum( ((xtest$observed - xtest$expected)^2)/xtest$expected )
```

```
## [1] 1.983758
```

Put that number into context?



χ^2 test

```
with(baseline, chisq.test(Sex, Genotype))
```

```
##
##     Pearson's Chi-squared test
##
## data:  Sex and Genotype
## X-squared = 1.9838, df = 2, p-value = 0.3709
```

Review

- We asked whether the sex ratio in the study was likely to be random, assuming an equal chance of an experimental mouse being male or female.
- We simulated 1000 studies under the assumption of n=266 and the odds of being female = 50%
- This is the null hypothesis.
- Our random data simulations test the null hypothesis: what would happen if we ran the experiment again and again and again under the same conditions assuming random assignment of males and females?
- Our p-value - the long run probability under repeated experiments - was vanishingly small.

So the choice of sex was almost certainly non-random. Did it matter?

- Chi-squared test to assess contingency tables.
- test outcome against known distribution
- the sex bias was shared by all genotypes - p value indicated this could easily have occurred by random chance.

Literate programming

Literate programming

The Idea:

- mix code, text, and figures in one document.
- All analyses and their outputs remain in sync
- Can work as a notebook

The Implementation:

- rmarkdown
    - simple markup language for text
    - code embedded in document
    - documents are compiled - or knitted - to produce output html or pdf
- Great alternative: Jupyter

Assignment

Group assignment number 1

1. Assemble into your assigned teams.
2. Ensure that RStudio is running and you can load all required libraries.
3. Load the required data
4. Create an rmarkdown document that contains the following:
    1. A summary table of the subject numbers per timepoint, genotype, and condition
    2. Visualization(s) of the difference in hippocampal volume by Genotype at the final timepoint.
    3. Visualization(s) of the difference in hippocampal volume by Condition at the final timepoint.
    4. Visualization(s) of the change over time by Condition and Genotype.
5. Make sure that all team members are listed as authors.
6. Any questions: ask here in person, or email us (jason.lerch@ndcn.ox.ac.uk, mehran.karimzadehreghbati@mail.utoronto.ca) and we promise to answer quickly.