IA1

Jake VanCampen

Oct 24 2017

```
knitr::opts_chunk$set(message=FALSE, warning = FALSE)
```

Enriched diet affects zebrafish size

To study the effects of an enriched diet on mean zebrafish size, 200 freshly hatched zebrafish were taken from the same clutch and randomly assigned to an enriched or a control diet (unenriched). The diet was administered for two months at which point the standard length (mm) and mass (kg) of each fish was measured. Results were compiled in a tab-separated file containing the following columns: "Individual", "Diet", "SL", and "Weight."

The following are the hypothesis to be tested in this experiment:

H_0: Feeding zeebrafish an enriched diet will not significantly affect their mean size after two months

H_A: Feeding zebrafish an enriched diet will significantly affect their mean size after two months

Exploratory Data Analysis

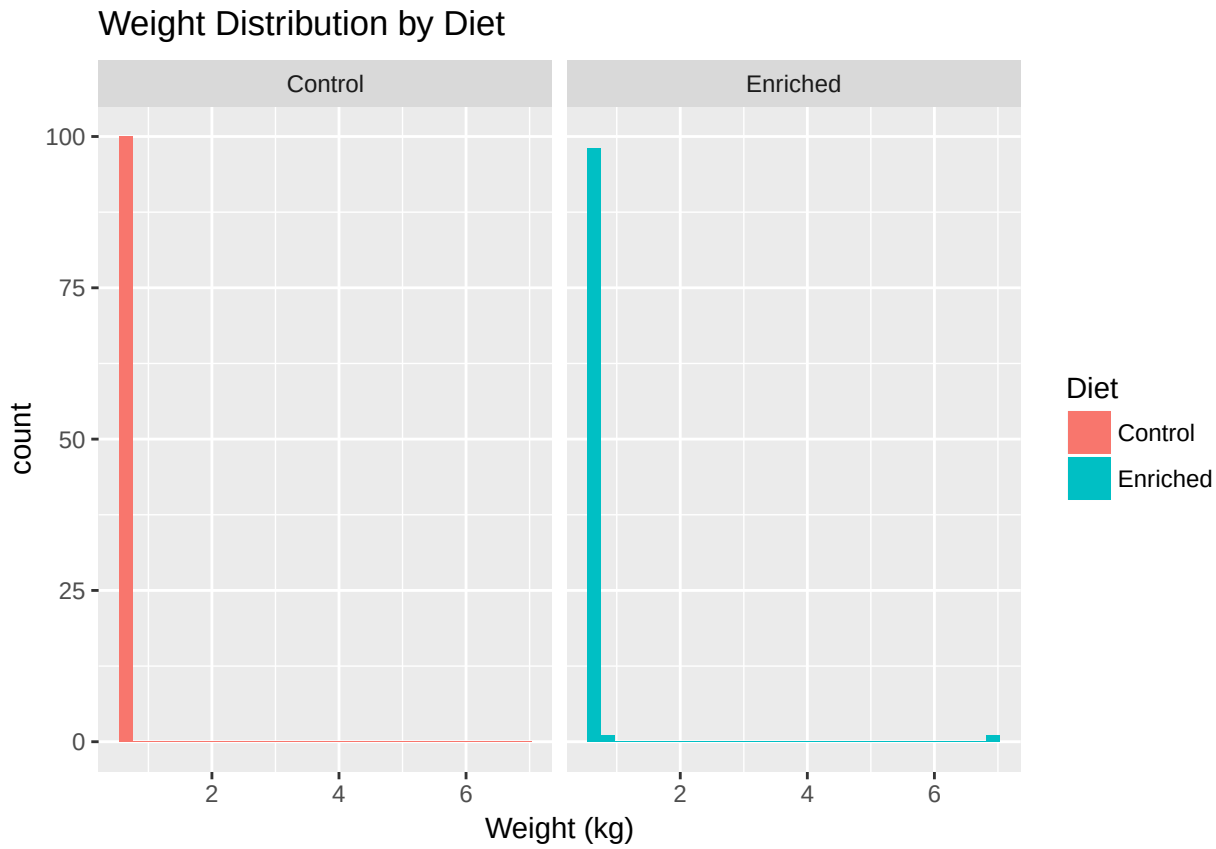
```
library(tidyverse)
library(magrittr)
library(ggpubr)
library(knitr)
library(lmodel2)
library(broom)
zfish_raw <- read_tsv('zfish_diet_IA.tsv')
head(zfish_raw)
```

```
## # A tibble: 6 x 4
##   Individual     Diet    SL Weight
##        <int>    <chr> <dbl>  <dbl>
## 1          1  Control  3.58  0.554
## 2          2 Enriched  4.66  0.699
## 3          3 Enriched  4.50  0.644
## 4          4  Control  3.71  0.624
## 5          5  Control  4.36  0.688
## 6          6  Control  3.89  0.613
```

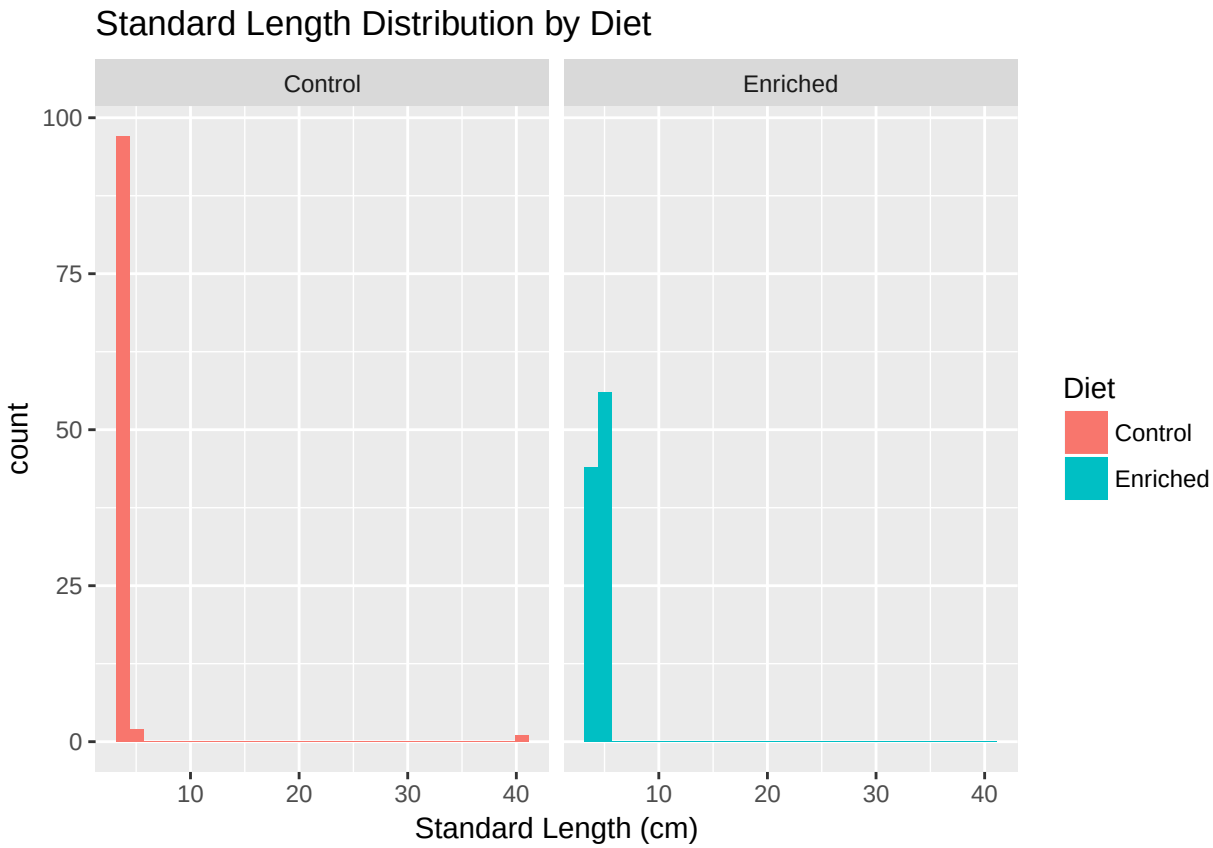
Histograms

Let's take a look at some of the ways the data are distributed. First we can look at histograms of continuous variables weight and standard length

```
ggplot(zfish_raw, aes(fill = Diet))+
  geom_histogram(aes(Weight))+
  facet_wrap(~Diet)+
  xlab('Weight (kg)')+
  ggtitle('Weight Distribution by Diet')
```



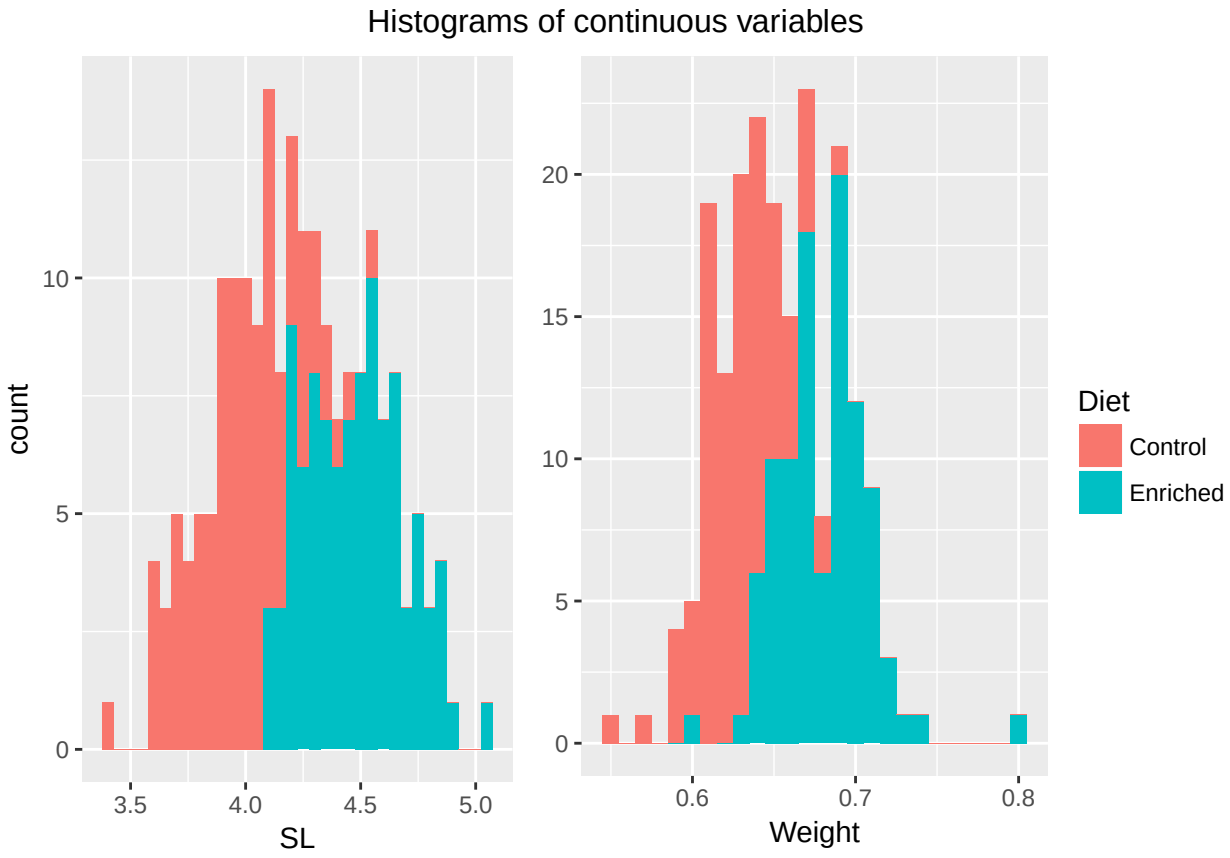

```
ggplot(zfish_raw, aes(fill = Diet))+
  geom_histogram(aes(SL))+
  facet_wrap(~Diet)+
  xlab('Standard Length (cm)') +
  ggtitle('Standard Length Distribution by Diet')
```



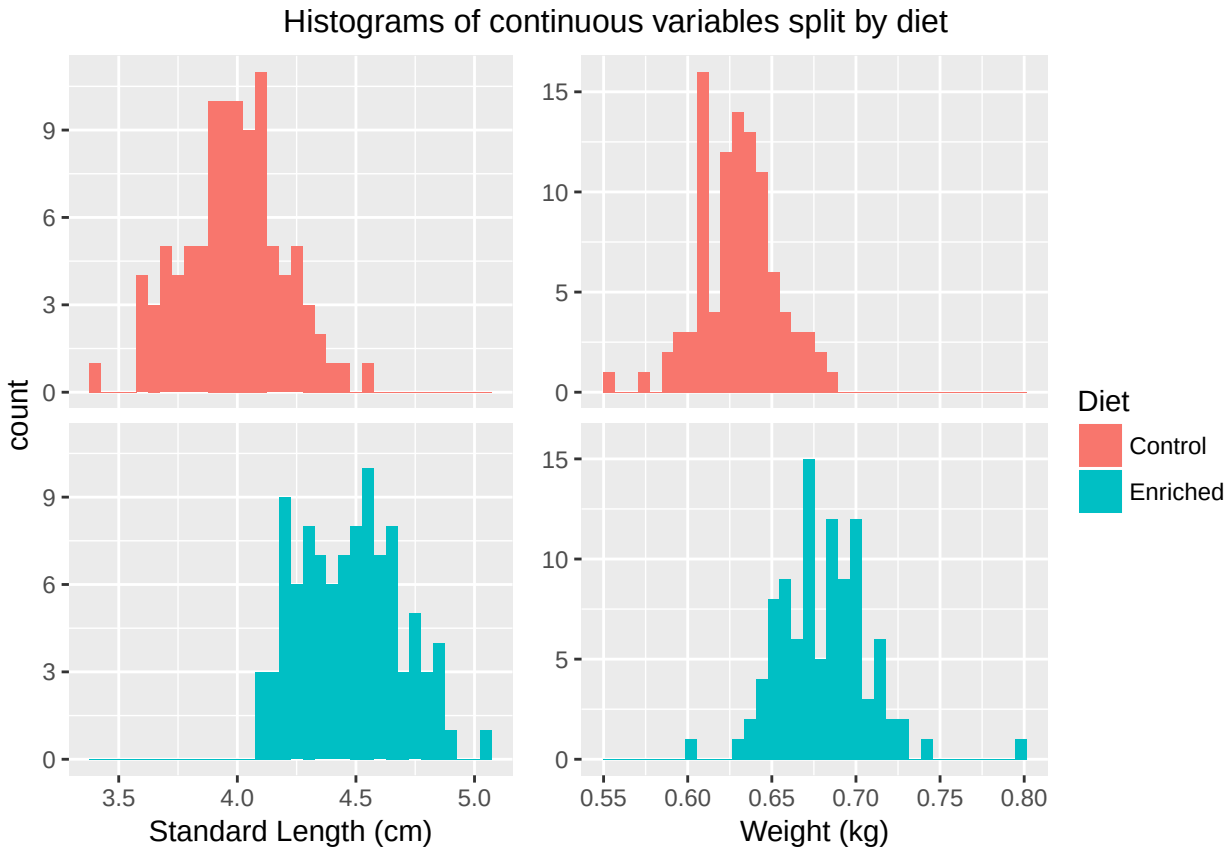
There appear to be erroneous data that is obscuring the distribution of both SL, and Weight. I will remove the obvious errors from the dataset as shown below:

```
zfish_correct <- zfish_raw %>%
  filter(., Weight < 4 & SL < 20)


# total SL hist colored by diet
dt_n <- ggplot(zfish_correct, aes(SL, fill = Diet))+
            geom_histogram(binwidth = 0.05)


# SL hist faceted by diet
dt <- ggplot(zfish_correct, aes(SL, fill = Diet)) +
          geom_histogram(binwidth = 0.05)+
          facet_grid(Diet~.)+
          xlab('Standard Length (cm)') +
  theme(
  strip.background = element_blank(),
  strip.text.x = element_blank(),
  strip.text.y = element_blank()
)


# total weight hist colored by diet
wt_n <- ggplot(zfish_correct, aes(Weight, fill = Diet)) +
            geom_histogram(binwidth = 0.01)+
```

```
            theme(axis.title.y = element_blank())

# weight hists faceted by diet
wt <- ggplot(zfish_correct, aes(Weight, fill = Diet)) +
  geom_histogram(aes(Weight, fill = Diet), binwidth = 0.007) +
  facet_grid(Diet~.)+
  xlab('Weight (kg)') +
  theme(
  strip.background = element_blank(),
  strip.text.x = element_blank(),
  strip.text.y = element_blank(),
  axis.title.y = element_blank()
)


# group figures
ggarrange(dt_n, wt_n, common.legend = TRUE, legend = 'right') %>% annotate_figure(., top = text_grob("Hi
```




```
ggarrange(dt, wt, common.legend = TRUE, legend = 'right') %>% annotate_figure(., top = text_grob("Histog
```

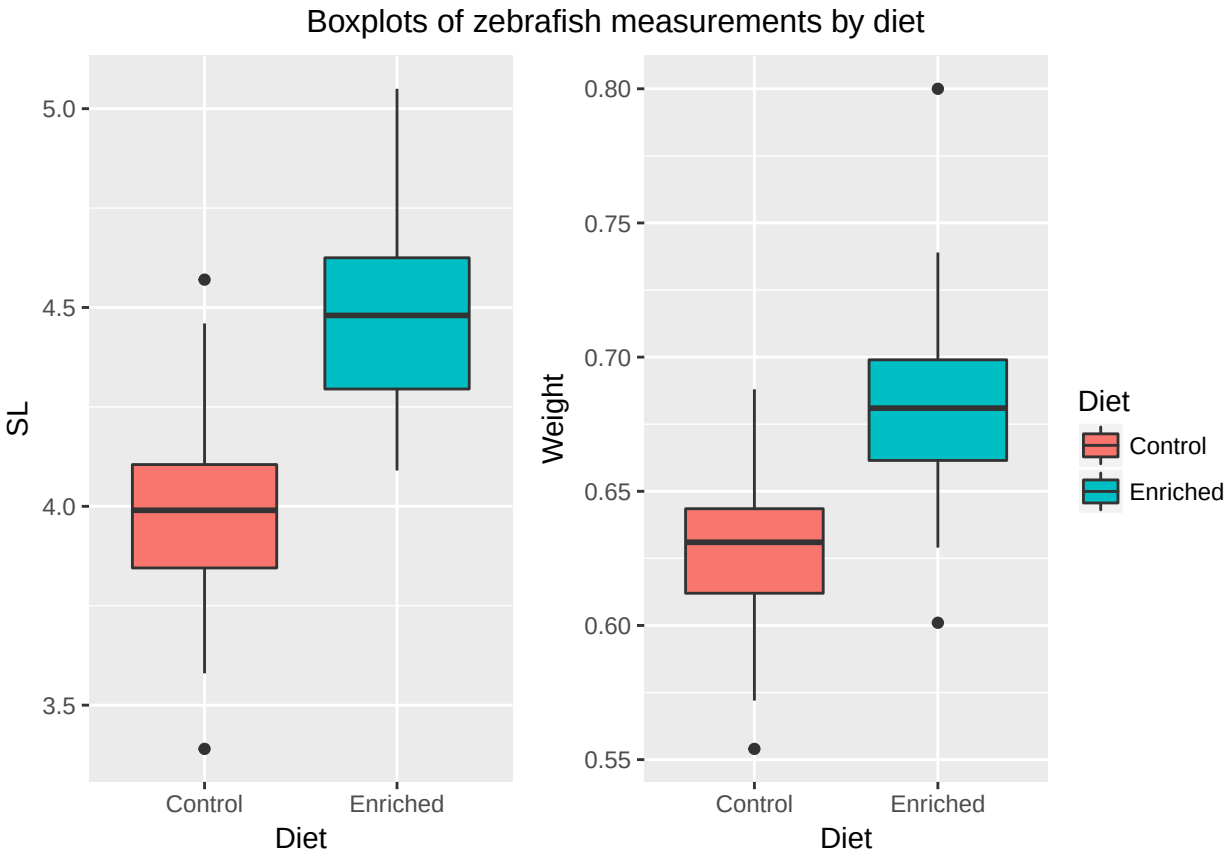


Okay that certainly changed the histograms! These data seem to be fairly normally distributed. The histograms of continuous variables split by diet show a positive shift in both continuous variables for the enriched diet compared to the control, let's do some statistics to back this up!

Boxplots

First similar comparisons will be made using boxplots.

```
bp_d <- ggplot(zfish_correct, aes(Diet, SL, fill = Diet))+
  geom_boxplot()

bp_w <- ggplot(zfish_correct, aes(Diet, Weight, fill = Diet))+
  geom_boxplot()

ggarrange(bp_d, bp_w, common.legend = TRUE, legend = 'right')  %>% annotate_figure(., top = text_grob("
```



Again we see the overall increase in both metrics with an enriched diet as compared to the control.

Summary statistics

And a summary table to compare relevant statistics:

```
Summ_stats <- zfish_correct %>%
  group_by(Diet) %>%
  summarise_at(vars(SL, Weight), c('mean', 'sd', 'var'))

kable(Summ_stats, align = 'c', caption = 'Summary statistics for zebra fish measuremnets by diet')
```

Table 1: Summary statistics for zebra fish measuremnets by diet

Diet	SL_mean	Weight_mean	SL_sd	Weight_sd	SL_var	Weight_var
Control	3.982323	0.6301717	0.2112201	0.0230740	0.0446139	0.0005324
Enriched	4.473434	0.6806566	0.2134085	0.0272105	0.0455432	0.0007404

T Test

Because the data look to be normally distributed, appear to have equal variances (var[control] ~= var[diet]), and occur as independant observations, the distributional assumptions are met to perform a parametric T Test. The T Test will test a difference in the means of the Control and Enriched treatments for both weight, and standard length.

```r
# T tests for each
sl_T <- t.test(zfish_correct$SL ~ zfish_correct$Diet)
wt_T <- t.test(zfish_correct$Weight ~ zfish_correct$Diet)


# format for reporting t test
t.report <- function(tt){
  tvalue <- tt$statistic %>% formatC(digits = 2, format = "f")
  pvalue <- tt$p.value %>% formatC(digits = 2, format = "E")
  if (round(tt$parameter, 0) == tt$parameter) {
    df <- tt$parameter
  } else {
    df <- formatC(tt$parameter, digits = 2, format = "f")
  }
  if (tt$p.value < 0.0005) {
    pvalue <- " < 0.001"
  } else {
    if (tt$p.value < 0.005) {
      pvalue <- paste0(" = ",tt$p.value %>% formatC(., digits = 3, format = "f"))
    } else {
      pvalue <- paste0(" = ",tt$p.value %>% formatC(., digits = 2, format = "f"))
    }
    }
  paste0("*t*(",df,") = ",tvalue, ", *p*", pvalue)
  }
```

Now the statement can be made that zebrafish have a significantly higher standard length when fed an enriched diet as compared to a normal diet, $t(195.98) = -16.27$, $p < 0.001$, and have a significantly higher weight when fed an enriched diet as compared to a normal diet, $t(190.90) = -14.08$, $p < 0.001$. The null hypothesis that there is no significant effect of an enriched diet on mean fish size can be rejected.

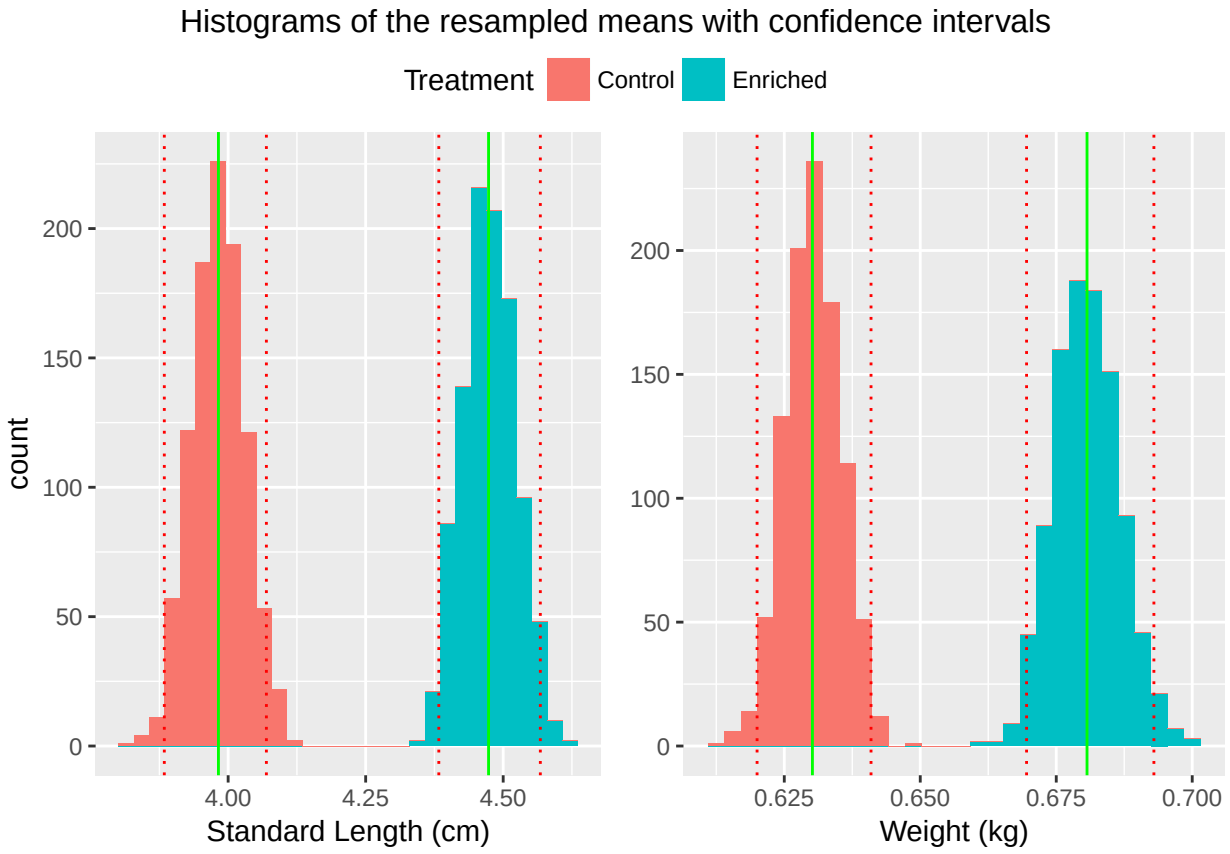
Non-parametric T test

The hypothesis testing can be done using a non-parametric parameter estimation (calculated by resampling of the means) and then performing a T-test. This resampling will cooerce the distribution of the parameter estimate into normality (if the data are not already normally distributed) before performing hypothesis tests, such that our confidence interval about the the hypothesis test is valid.

```r
# bootstrap_means func
boot <- function(x) {
z <- NULL
for (i in 1:1000) {
  xboot <- sample(x, 20, replace = T)
  z[i] <- mean(xboot)
}
Mean <- mean(z)
SEM <- sd(z)
CI <- quantile(z,c(0.025,0.975))
return(list(tbl_df(data.frame('Mean' = Mean,
                  'SEM' = SEM,
                  bm.lower.ci = CI[1],
                  bm.upper.ci = CI[2])), tbl_df(data.frame(z))))
}
```

```r
Control <- zfish_correct %>% filter(., Diet == 'Control')
Enriched <- zfish_correct %>% filter(., Diet == 'Enriched')

# Standard length stats and distribtion
sl_bootC_st <- boot(Control$SL)[1]
sl_bootC_dist <- boot(Control$SL)[2]

sl_bootE_st <- boot(Enriched$SL)[1]
sl_bootE_dist <- boot(Enriched$SL)[2]

# combine
sl_bootC <- tbl_df(data.frame(sl_bootC_dist, rep('Control', 1000)))
colnames(sl_bootC) <-  c('boot_means_sl', 'Treatment')


sl_bootE <- tbl_df(data.frame(sl_bootE_dist, rep('Enriched', 1000)))
colnames(sl_bootE) <-  c('boot_means_sl', 'Treatment')

# stack
sl_boot <- rbind(sl_bootC, sl_bootE)

# Weight stats and distribution
wt_bootC_st <- boot(Control$Weight)[1]
wt_bootC_dist <- boot(Control$Weight)[2]

wt_bootE_st <- boot(Enriched$Weight)[1]
wt_bootE_dist <- boot(Enriched$Weight)[2]

#combine
wt_bootC <- tbl_df(data.frame(wt_bootC_dist, rep('Control', 1000)))
colnames(wt_bootC) <-  c('boot_means_wt', 'Treatment')

wt_bootE <- tbl_df(data.frame(wt_bootE_dist, rep('Enriched', 1000)))
colnames(wt_bootE) <-  c('boot_means_wt', 'Treatment')


# stack
wt_boot <- rbind(wt_bootC, wt_bootE)


# plot resamples means for standard length
resampled_sl <- ggplot(sl_boot, aes(boot_means_sl, fill = Treatment))+
  geom_histogram()+
  xlab('Standard Length (cm)')+
  geom_vline(xintercept = sl_bootE_st[[1]]$bm.lower.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = sl_bootE_st[[1]]$bm.upper.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = sl_bootC_st[[1]]$bm.lower.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = sl_bootC_st[[1]]$bm.upper.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = Summ_stats$SL_mean, color = 'green')



resampled_wt <- ggplot(wt_boot, aes(boot_means_wt, fill = Treatment))+
```

```
  geom_histogram()+
  xlab('Weight (kg)')+
  geom_vline(xintercept = wt_bootE_st[[1]]$bm.lower.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = wt_bootE_st[[1]]$bm.upper.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = wt_bootC_st[[1]]$bm.lower.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = wt_bootC_st[[1]]$bm.upper.ci, color = 'red', linetype="dotted") +
  geom_vline(xintercept = Summ_stats$Weight_mean, color = 'green') +
  theme(
    axis.title.y = element_blank()
  )



ggarrange(resampled_sl, resampled_wt, common.legend = TRUE) %>% annotate_figure(., top = text_grob("His
```



SL AND WEIGHT

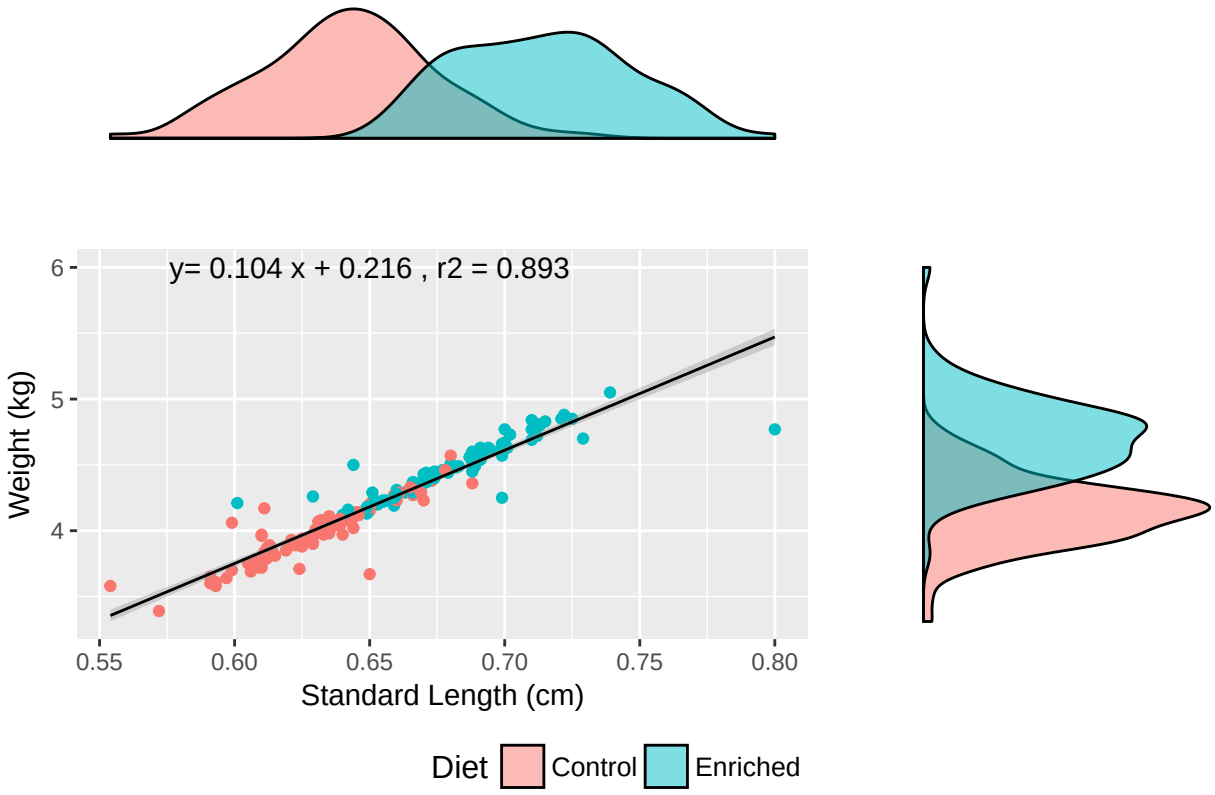
```
#type I model fit
zfish_lm <- lm(zfish_correct$Weight~zfish_correct$SL)
b <- format(data.frame(tidy(zfish_lm))$estimate[1],digit = 3)
m <- format(data.frame(tidy(zfish_lm))$estimate[2],digits = 3)
glance(zfish_lm)
```

```
##   r.squared adj.r.squared      sigma statistic      p.value df   logLik
## 1 0.8934216     0.8928779 0.01168033  1643.022 3.097717e-97  2 601.1253
```

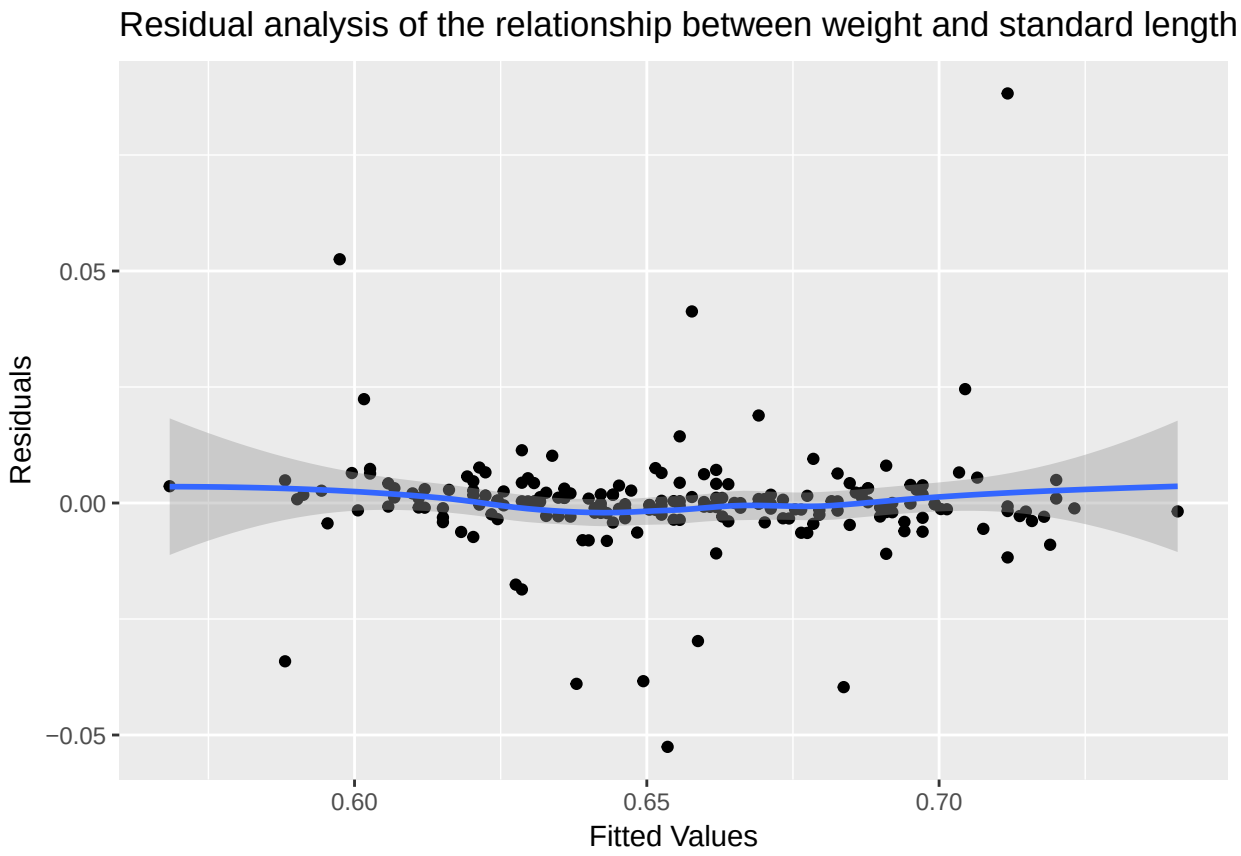
```
##         AIC       BIC   deviance df.residual
## 1 -1196.251 -1186.386 0.02674031         196
```

```
r2 <- format(data.frame(glance(zfish_lm)$r.squared)[[1]], digits = 3)

eq <- paste("y=", m, "x +", b,',', 'r2', "=", r2)


# plot the relationship between standard lenght and weight
corr <- ggplot(zfish_correct, aes(zfish_correct$Weight,
                          zfish_correct$SL,
                          color = Diet)) +
  geom_point() +
  geom_smooth(method = 'lm', color = 'black', size = 0.5) +
  xlab('Standard Length (cm)') +
  ylab('Weight (kg)')+
  annotate('text',x = 0.65, y = 6, label = eq)

sl_dens <- ggdensity(zfish_correct, 'SL', fill = 'Diet') + clean_theme()

wt_dens <- ggdensity(zfish_correct, 'Weight', fill = 'Diet') + rotate() + clean_theme()

# Arranging the plot
ggarrange(sl_dens, NULL, corr, wt_dens,
          ncol = 2, nrow = 2,  align = "hv",
          widths = c(2, 1), heights = c(1, 2),
          common.legend = TRUE, legend = 'bottom') %>% annotate_figure(top = text_grob('The relationshi
```

The relationship between Weight and Standard length over a change in diet



```r
# combine residuals vs. fitted values
res <- tbl_df(data.frame(cbind(zfish_lm$residuals, zfish_lm$fitted.values)))

ggplot(res, aes(X2,X1)) +
  geom_point() +
  geom_smooth(mode = 'lm') +
  xlab('Fitted Values')+
  ylab('Residuals')+
  ggtitle('Residual analysis of the relationship between weight and standard length')
```



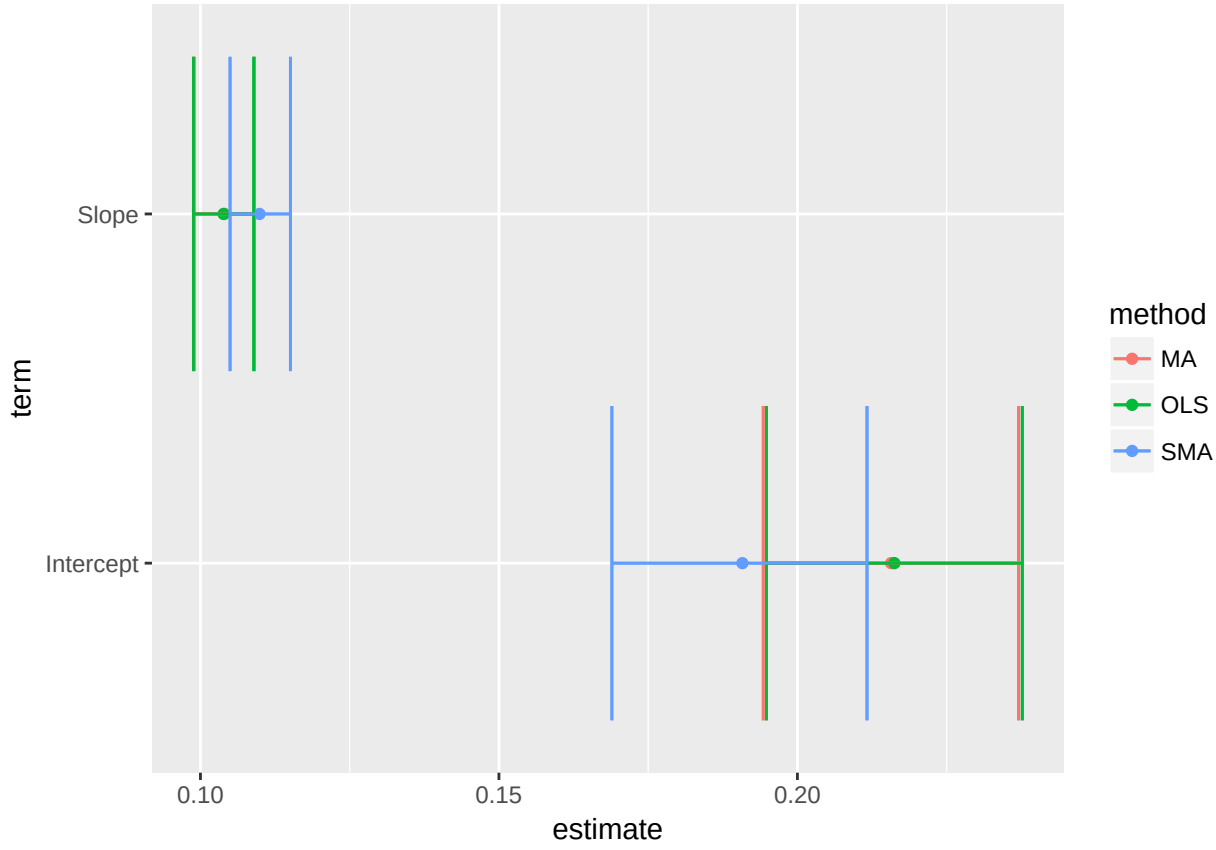

```r
# type II model fit
zfish_lm2 <- lmodel2(zfish_correct$Weight~zfish_correct$SL)
tidy(zfish_lm2)
```

```
##   method      term  estimate   conf.low conf.high
## 1     MA Intercept 0.2156807 0.19427461 0.2370645
## 2     MA     Slope 0.1040081 0.09895024 0.1090711
## 3    OLS Intercept 0.2162407 0.19481062 0.2376707
## 4    OLS     Slope 0.1038756 0.09882167 0.1089295
## 5    SMA Intercept 0.1907835 0.16892498 0.2116598
## 6    SMA     Slope 0.1098969 0.10495909 0.1150670
```

```r
glance(zfish_lm2)
```

```
##   r.squared    theta      p.value           H
## 1 0.8934216 0.701477 3.097717e-97 2.50623e-05
```

```
ggplot(tidy(zfish_lm2), aes(estimate, term, color = method)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high))
```




```
  tidy(zfish_lm2)
```

```
##   method      term  estimate   conf.low conf.high
## 1     MA Intercept 0.2156807 0.19427461 0.2370645
## 2     MA     Slope 0.1040081 0.09895024 0.1090711
## 3    OLS Intercept 0.2162407 0.19481062 0.2376707
## 4    OLS     Slope 0.1038756 0.09882167 0.1089295
## 5    SMA Intercept 0.1907835 0.16892498 0.2116598
## 6    SMA     Slope 0.1098969 0.10495909 0.1150670
```

There is a significant relationship between Weight(kg) and Standard Length (cm) $F1{,}196 = 1643$, $p < 0.001$