

Assignment_6

Mingrui Duan

```
# Load the package essential
library(tidyr)
library(dplyr)
library(readr)
```

Question 1: Load the data and convert it to a format that ggplot will like. In doing this, you should make a new data frame containing the “famtot” and “unfamtot” times with an identifier for each subject and a column for condition. Be sure you only use the data from Experiment 5. Note that you should divide famtot and unfamtot by 30 to convert to seconds. After you make it, print a summary of this data frame.

```
# Load the original data
d <- read.csv("LUL_pss.csv")
# Get the new raw dataframe with the columns we want
d <- subset(d, select = c("exp5", "famtot", "unfamtot"))
# Clean all the NA values
d <- d[complete.cases(d),]
# Divide the value by 30
d$famtot <- d$famtot/30
d$unfamtot <- d$unfamtot/30

# Compute each condition's mean & sd
famtot.mean <- mean(d$famtot)
famtot.sd <- sd(d$famtot)
unfamtot.mean <- mean(d$unfamtot)
unfamtot.sd <- sd(d$unfamtot)

# Reshape the raw dataframe
d <- d %>%
  pivot_longer(
    cols = ends_with("tot"),
    names_to = "condition",
    values_to = "total_gaze",
    values_drop_na = TRUE
  )

d <- subset(d, select = c("condition", "total_gaze"))

# Print a summary of that dataset
summary(d)
```

```
##      condition      total_gaze
```

```
## Length:46      Min.   :23.17
## Class :character 1st Qu.:27.92
## Mode  :character Median :31.08
##              Mean  :30.92
##              3rd Qu.:33.54
##              Max.   :38.90
```

Question 2: Since T-tests assume normality, first use `geom_qq` to check this assumption. Make a QQ-plot that is faceted by condition. Does yours match how a QQ-plot should look for normal data, or not? Explain in 1-2 sentences.

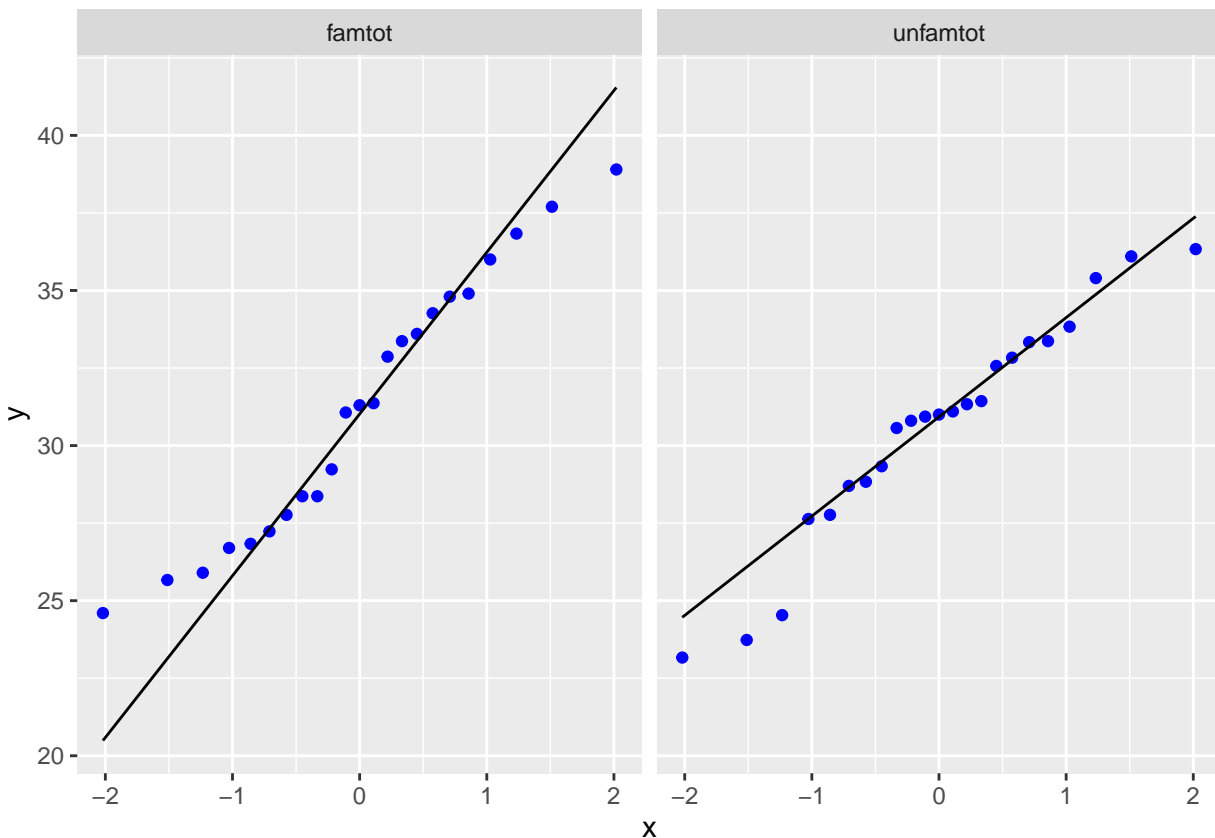
```
library("ggplot2")
set.seed(123)
# Create the theoretical normal distribution of two conditions
famtot.norm <- rnorm(23, mean = famtot.mean, sd = famtot.sd)
unfamtot.norm <- rnorm(23, mean = unfamtot.mean, sd = unfamtot.sd)

# Adjust the order in condition column
d$condition <- c(subset(d, condition == "famtot")$condition,
                 subset(d, condition == "unfamtot")$condition)

d$sample <- c(famtot.norm, unfamtot.norm)

# sample_data <- data.frame(
#   c(famtot.norm, unfamtot.norm),
#   condition = c(rep("famtot", 23), rep("unfamtot", 23)))

ggplot(d, aes(sample = total_gaze)) +
  geom_qq(color = "blue") +
  geom_qq_line() +
  facet_wrap(~ condition)
```



I would say my qq plot is not that matched because some points have significant away from the refer line. The reason is that if the data is the normal data, then the points roughly follow a straight line. But I observed some curves in it.

Question 3: Read about the Shapiro-Wilk test online. This is commonly used as a test for nonnormality. (a) What is the null hypothesis? What's the alternative to the null hypothesis? (b) Run the test in R (you can search for how) and report the results. (c) In 1-2 sentences, explain what the test would typically look like if the data was normal, vs. if it was not normal. (d) In 1-2 sentences, explain what the p-value you got means.

- (a) The null hypothesis is that the data is normally distributed; While the alternative is that the data is not normally distributed.
- (b) Run the test:

```
library("ggplot2")
library("stats")

gaze.fam <- subset(d, condition == "famtot")$total_gaze
gaze.unfam <- subset(d, condition == "unfamtot")$total_gaze

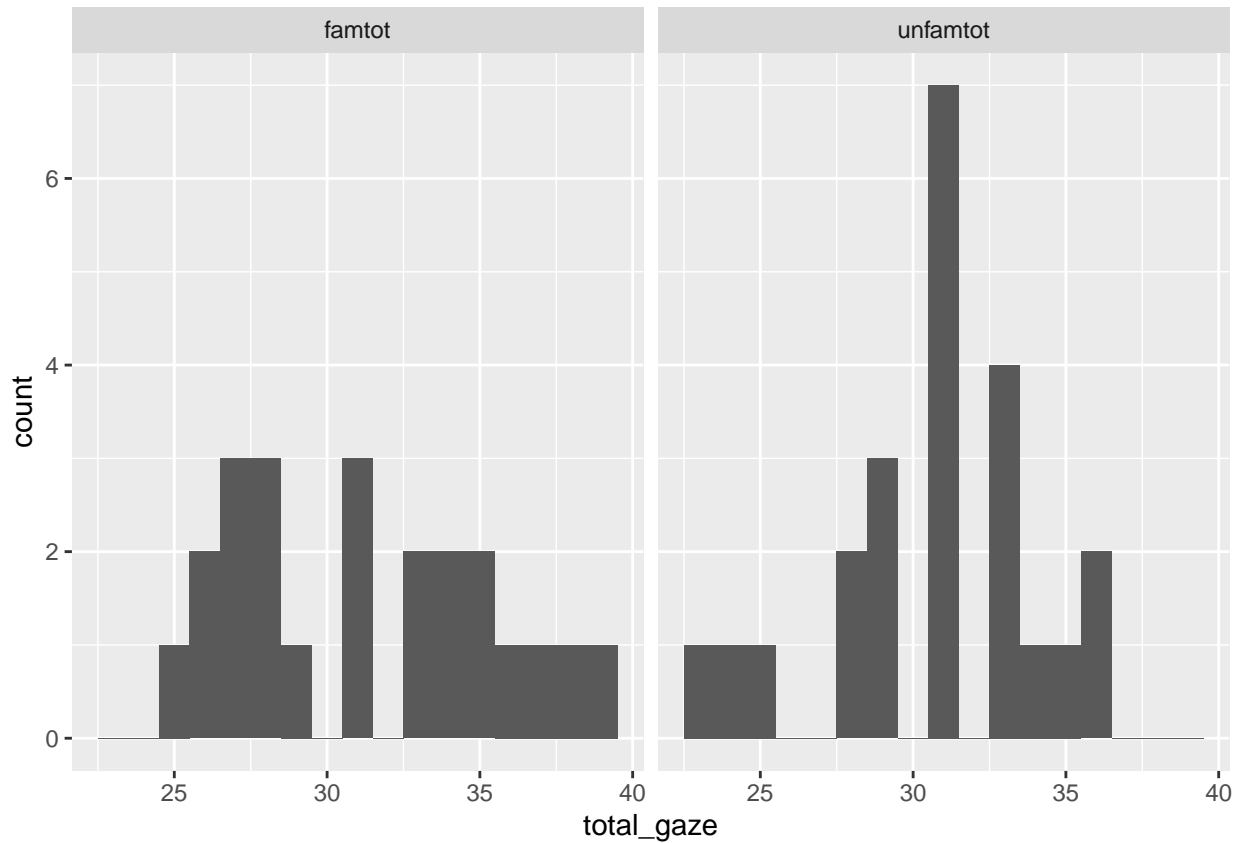
# First test the data based on different condition
shapiro.test(gaze.fam)
```

```
##
## Shapiro-Wilk normality test
##
## data: gaze.fam
## W = 0.95372, p-value = 0.3488
```

```
shapiro.test(gaze.unfam)
```

```
##
## Shapiro-Wilk normality test
##
## data: gaze.unfam
## W = 0.95199, p-value = 0.3217
```

```
# Visualize it
ggplot(d, aes(x = total_gaze)) +
  geom_histogram(binwidth = 1) +
  # annotate("text",
  #         x = Inf,
  #         y = Inf,
  #         label = paste0("p-value = ",
  #                         round(shapiro.test(d$total_gaze)$p.value, 3)))
  facet_wrap(~ condition)
```



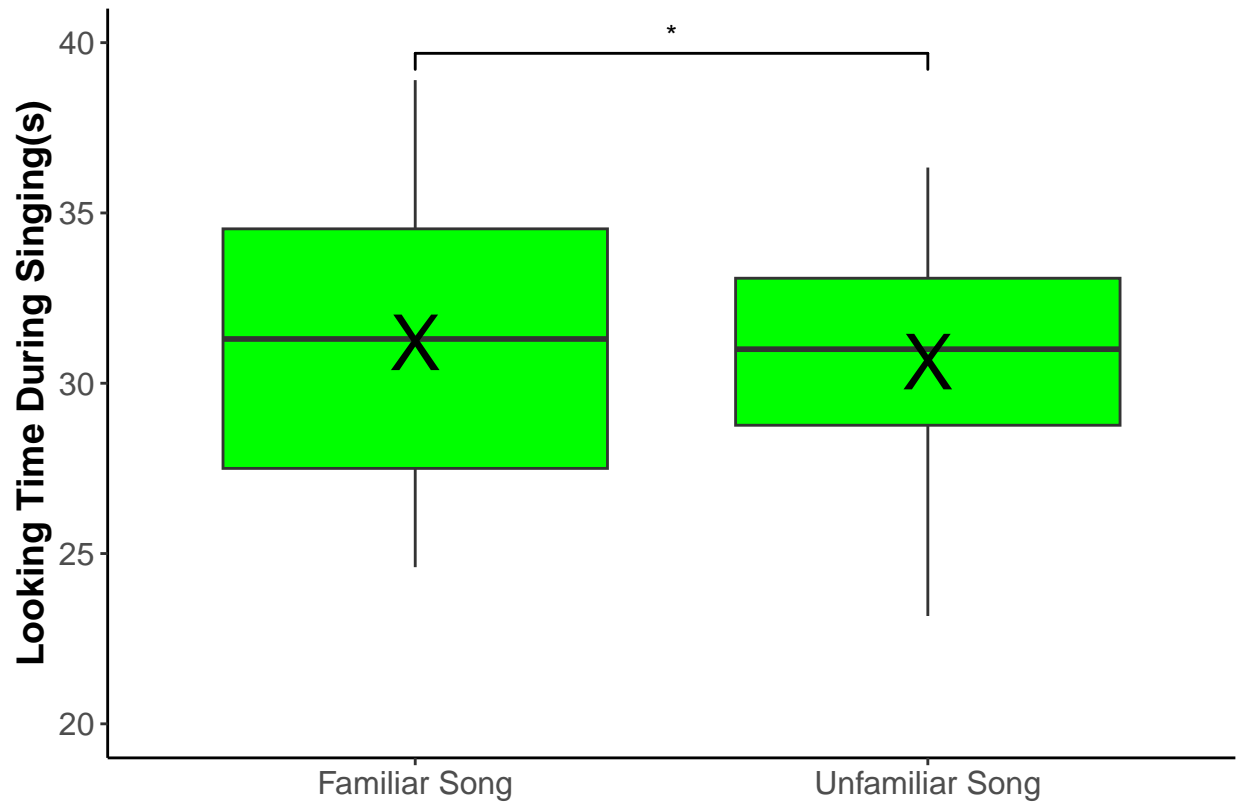
- (c) If the data was normal, the test should be like having a high p-value also visually this would be reflected in a histogram of the data that is bell-shaped and symmetric; If the data was not normal, the test should be like having a low p-value also visually this would be reflected in a histogram of the data that is skewed or has a non-symmetric shape.
- (d) The p-value I got from the test means that whether the null hypothesis is rejected depending on the chosen significance level. My null hypothesis will be rejected because my p-value is low so that there is no strong evidence to reject the null hypothesis of normality.

Question 4: Replicate Figure 6b from this data. Be sure to get (a) the blank background, (b) the green boxes, (c) the “X”es on the means, (d) the axis labels and ranges, and (e) the significance lines. For these significance markers, (e), you should install the package “ggsignif” and read its help file to see how it works. Note that this is a SOLO question, but you may ask questions about ggsignif after reading its help files. You will need to provide `annotation="*"` to `geom_signif` (under Advanced Example) rather than relying on its default. For any of the other options, please google and read the help files; if you can’t get it quite perfect, don’t worry. Get it as close as you can.

```
library("ggsignif")

ggplot(d, aes(x = condition, y = total_gaze)) +
  geom_boxplot(fill = "green") +
  stat_summary(fun.y = mean,
              geom = "point",
              shape = 'X', size = 10) +
  geom_signif(comparisons = list(c("famtot", "unfamtot")),
              annotations = "*") +
  scale_x_discrete(labels = c("Familiar Song", "Unfamiliar Song")) +
  xlab("") +
  ylim(c(20, 40)) +
  ylab("Looking Time During Singing(s)") +
  theme(panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14, face = "bold"))
```

```
## Warning: The 'fun.y' argument of 'stat_summary()' is deprecated as of ggplot2 3.3.0.
## i Please use the 'fun' argument instead.
```



Question 5: (a) Is a paired or an unpaired t-test the most appropriate to use on this data? Write 2-3 sentences like you might write in a paper's methods section to justify your answer. (b) Run whichever is right and report your results. (c) Run the other one (that you don't think is right to use) and compare your answers. Are they the same or different? Why? (d) Write 2 sentences politely responding to a hypothetical reviewer who says you should run a single-sample t-test on the difference between conditions for each subject

(a) I think unpaired t-test is more appropriate in this data. Because unpaired t-test is used when two groups being compared are independent of each other. In this data, the total gaze time of each infant doesn't affect other's gaze time also my data follow the normal distribution.

(b)

```
# Conduct unpaired t-test
t.test(gaze.fam, gaze.unfam)

##
## Welch Two Sample t-test
##
## data: gaze.fam and gaze.unfam
## t = 0.48706, df = 42.923, p-value = 0.6287
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.775211 2.905646
## sample estimates:
## mean of x mean of y
## 31.20145 30.63623
```

(c)

```
# Conduct paired t-test
t.test(gaze.fam, gaze.unfam, paired = TRUE)

##
## Paired t-test
##
## data: gaze.fam and gaze.unfam
## t = 0.48037, df = 22, p-value = 0.6357
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
## -1.874969 3.005404
## sample estimates:
## mean difference
## 0.5652174
```

(d) The data are not independent because each subject contributes two data points (one for each condition). Therefore, using a single-sample t-test would violate the assumption of independence and the resulting p-values would be inaccurate.

Question 6: Run a `wilcox.test` (you decide whether its paired or unpaired) on the differences. Do your results agree with Q5 or not? If not, write 2-3 sentences explaining why.

```
wilcox.test(gaze.fam, gaze.unfam)
```

```
## Warning in wilcox.test.default(gaze.fam, gaze.unfam): cannot compute exact  
## p-value with ties
```

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data: gaze.fam and gaze.unfam  
## W = 283, p-value = 0.6925  
## alternative hypothesis: true location shift is not equal to 0
```

My results agree with the Q5. Because here I got p-value which is 0.6925, which means I will accept the null hypothesis that there is no difference between two groups. Back in my t-test in Q5, we got the same conclusion because our p-value in t-test is also have no strong confidence to reject the null hypothesis.

Question 7: If the t-test shows a statistically significant difference but the wilcox test did not, would it be ethical to publish? Explain why in 2-3 sentences. Is it ethical to always check if either is significant and publish if it is? Write 2-3 sentences on why.

1. It is ethical to publish because here our sample size is small(23). Also, I think the whole research is well-executed and the design is sound.
2. It is ethical to always check if either is significant and publish if it is. Cause in different studies we have different factors and design. So sometimes we have to balance all these factors and make final decision that whether to publish.