# Exploring the latency data

This Rmd file is where I try to figure out what's going on in the data.

# Load the latency datasetwd("../")source("R/01\_functions.r", chdir = TRUE)load("data/results.RData")

## Reponses to JE's email:

### *For methods, can you write bullet points on how we calculated latencies?*

Reaction times measure the latency between looking to the distractor image and shifting the gaze towards the target image after the onset of the target word. The following conditions were required for the latency calculation:

* During the first 50 ms of the target word, the child had to be looking onscreen, but not at the target image.
* The first look to the target must occur after 250 ms. (That is, shifts of looks towards the target before 250 ms were considered too fast to be deliberate responses to the target word.)

### *After trimming, what % of trials had latencies in CS1? In CS2? Are there more trials with latencies in CS2 compared to CS1 after trimming, as there were before trimming?*

Here are the summary stats for the unadjusted values. First, we remove subjects who have been designated as non-keepers. We also remove subjects who were tested with the AAE dialect stimuli, since those audio stimuli were not the same duration.

results <- subset(results, is.na(Keeper) & Dialect == "SAE")PrintDescriptives(results)

|  |  |  |
| --- | --- | --- |
|  | **CS1** | **CS2** |
| n | 285 | 579 |
| mean | 756.8 | 752.1 |
| sd | 451.1 | 484.7 |
| median | 699.5 | 616.2 |
| trimmed | 702.5 | 687.8 |
| mad | 345.7 | 370.4 |
| min | 66.62 | 66.62 |
| max | 2498 | 2482 |
| range | 2432 | 2415 |
| se | 26.72 | 20.14 |

Descriptives for unadjusted latency values

ComputePercentNA(results)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Version** | **NA Latencies** | **Real Latencies** | **Percent NA** |
| 1 | CS1 | 567.00 | 285.00 | 66.55 |
| 2 | CS2 | 345.00 | 579.00 | 37.34 |

Response rates for the experiment versions, before trimming.

Now we trim of the too-fast values using 250 ms as the cut-off.

results <- TrimTooFast(results, cutoff = 250)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Version** | **Num > 250 ms** | **Num < 250 ms** | **Num NA** |
| 1 | CS1 | 259.00 | 26.00 | 567.00 |
| 2 | CS2 | 531.00 | 48.00 | 345.00 |

The upper-bound of the trimming depends on what pool of latencies are used to compute the standard deviation used for the 2-SD cut-off.

ComputeUpperBound <- function(x) mean(x, na.rm = T) + (2 \* sd(x, na.rm = T))DropAboveUpperBound <- function(df) { cutoff <- ComputeUpperBound(df$Latency) df$Latency[df$Latency > cutoff] <- NA df}# Pooling both experiments togetherComputeUpperBound(results$Latency)

## [1] 1723

# Separating the two experimentsby(results$Latency, results$Version, ComputeUpperBound)

## results$Version: CS1## [1] 1674## -------------------------------------------------------- ## results$Version: CS2## [1] 1746

cs1 <- subset(results, Version == "CS1")cs1 <- DropAboveUpperBound(cs1)cs2 <- subset(results, Version == "CS2")cs2 <- DropAboveUpperBound(cs2)results <- rbind(cs1, cs2)

PrintDescriptives(results)

|  |  |  |
| --- | --- | --- |
|  | **CS1** | **CS2** |
| n | 243 | 496 |
| mean | 734.4 | 720.4 |
| sd | 285.2 | 350.5 |
| median | 699.5 | 632.9 |
| trimmed | 715.8 | 681.5 |
| mad | 296.3 | 345.7 |
| min | 249.8 | 249.8 |
| max | 1649 | 1715 |
| range | 1399 | 1466 |
| se | 18.29 | 15.74 |

Descriptives for trimmed latency values. Upper bounds were trimmed within each experimnet.

ComputePercentNA(results)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Version** | **NA Latencies** | **Real Latencies** | **Percent NA** |
| 1 | CS1 | 609.00 | 243.00 | 71.48 |
| 2 | CS2 | 428.00 | 496.00 | 46.32 |

Response rates for the experiment versions, after trimming.

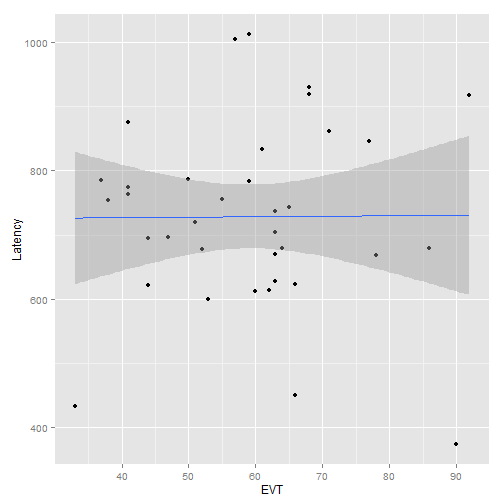
### *After trimming, is the average latency shorter for CS2 as compared to CS1? (as it was before trimming)?*

Yes.

### Is there a relationship between vocab size (either EVT-2 raw score of PPVT-4 raw score) and latency for CS1?

I'm going to compute the average latency within each subject and plot latency as a function of EVT and PPVT.

subject\_means <- ddply(results, "Subject", summarize, Version = unique(Version), EVT = unique(EVT), PPVT = unique(PPVT), Latency = mean(Latency, na.rm = TRUE))cs1 <- subset(subject\_means, Version == "CS1")# EVTqplot(data = cs1, x = EVT, y = Latency) + geom\_smooth(method = "lm")



plot of chunk unnamed-chunk-9

# PPVTqplot(data = cs1, x = PPVT, y = Latency) + geom\_smooth(method = "lm")

## Warning: Removed 2 rows containing missing values (stat\_smooth).

## Warning: Removed 2 rows containing missing values (geom\_point).

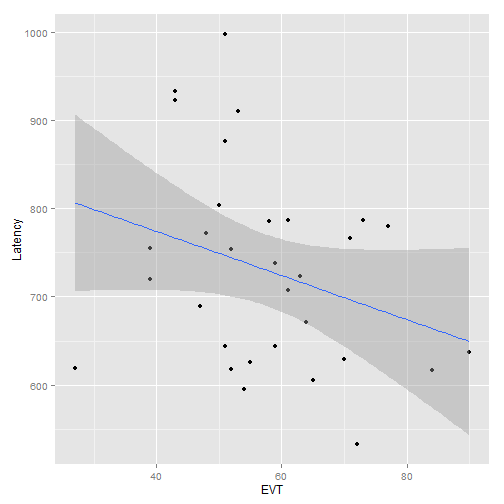


plot of chunk unnamed-chunk-9

1. Same question for CS2?

For the aggregated values, there is some correlation between EVT and latency.

cs2 <- subset(subject\_means, Version == "CS2")# EVTqplot(data = cs2, x = EVT, y = Latency) + geom\_smooth(method = "lm")

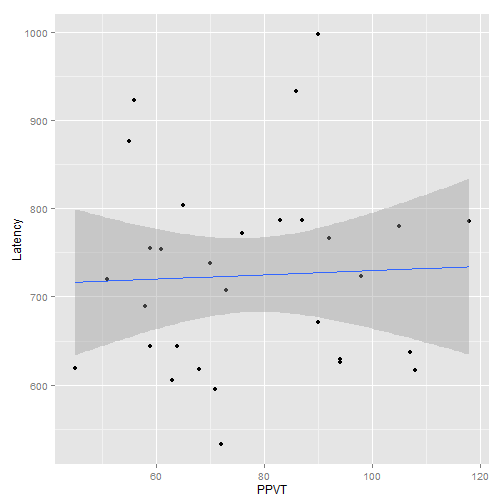


plot of chunk unnamed-chunk-10

# PPVTqplot(data = cs2, x = PPVT, y = Latency) + geom\_smooth(method = "lm")

## Warning: Removed 1 rows containing missing values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).



plot of chunk unnamed-chunk-10

summary(lm(Latency ~ EVT, cs2))

## ## Call:## lm(formula = Latency ~ EVT, data = cs2)## ## Residuals:## Min 1Q Median 3Q Max ## -188.4 -76.8 -12.4 67.2 251.1 ## ## Coefficients:## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 874.26 86.99 10.1 5.9e-11 \*\*\*## EVT -2.50 1.47 -1.7 0.1 ## ---## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1## ## Residual standard error: 110 on 29 degrees of freedom## Multiple R-squared: 0.0904, Adjusted R-squared: 0.059 ## F-statistic: 2.88 on 1 and 29 DF, p-value: 0.1