# Considering the Speed of Requested Books, and Answering the Question: How Long Will It Take?

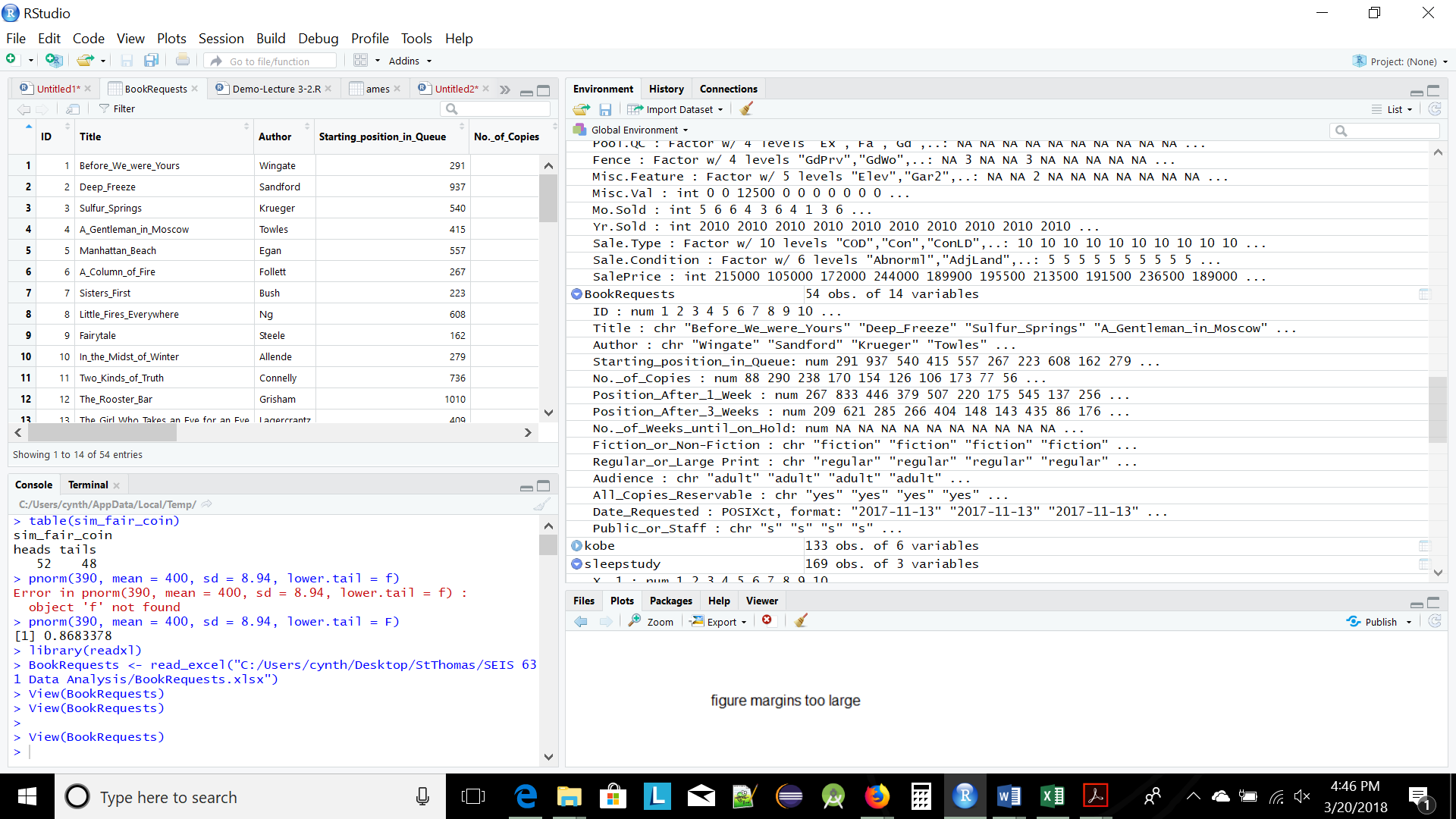
This data is from book requests at a library. From November 2017 (and is ongoing), I kept note of the date when the books were requested, how many copies of the title, it’s original position in the queue, and measured how quickly it moved through the queue. I obtained this information using an unused, personal account and requested books that I noticed to be popular, where people would ask how long it would take. I kept track of this information with print-outs, and screenshotting the account pages.

Each observation represents a requested book.

Table 1: Variable Names and Information

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| VARIABLE NAME | VARIABLE TYPE | VARIABLE DESCRIPTION | RANGE OR POSSIBLE VALUES | MISSING OR WEIRD VALUES? |
| ID | label | Identification of books, rows | 1-54 | no |
| Title | label | Title to identify book | N/A | no |
| Author | label | Author to identify book | N/A | no |
| Starting\_position\_  in\_Queue | Numeric discrete | This is the position in the queue when first requested. | 1 - 1066 | no |
| No.\_of\_Copies | Numeric discrete | The number of copies of book that is requested. | 1 - 292 | Yes  Depends on which catalog used. |
| Position\_After\_  1\_Week | Numeric discrete | The requests’ position in queue after a week. | 1 - 1104 | Yes  probably from cached values |
| Position\_After\_  3\_Weeks | Numeric discrete | The requests’ position in queue after 3 weeks. | 1 - 1000 | Yes  probably from cached values |
| Days\_until\_  on\_Hold | Numeric discrete | The number of days until it is on the hold shelf. | 1 - 100 | no |
| Fiction\_or\_Non-Fiction | Binary categorical | Whether the book is fiction or non-fiction. | Fiction or non-fiction | no |
| Regular\_or\_Large Print | Binary categorical | Whether the book is regular print or large print. | Regular print or large print | no |
| Audience | Categorical with three types | Whether the book is intended for adults, teens, or children. | Adults, teens, or children | no |
| All\_Copies\_  Reservable | Binary categorical | Whether all books are available to reserve. | Yes or no | no |
| Date\_Requested | Date | The date the request was made. | 11-13-2017 – 2-26-2018 | no |
| Public\_or\_Staff | Binary categorical | Was the information obtained from the staff or public catalog. | P or s | Yes  Assuming staff was live, and public was cached. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NEW VARIABLES | VARIABLE TYPE | VARIABLE DESCRIPTION | RANGE OR POSSIBLE VALUES | MISSING OR WEIRD VALUES? |
| Audience\_refined | Categorical with three types | Combined teen and children, split adult into fiction and non-fiction | Adult fiction, adult non-fiction, non-adult | No, but still too little data for non-adult |
| q\_start\_three\_wks | Numeric continuous | Measures speed from start to three weeks, daily movement | 0 - 1 | no |
| q\_daily | Numeric continuous | Movement of a book through the queue per day | 0 - 1 | NA, some books have not been completed |
| Is\_popular | Binary categorical | Whether more popular than the mean | T or F | no |
| extremes | Binary categorical | Whether they contain an extreme value of number of copies | T or F | no |



# Section 2

**Some Basic Data Checks:**

The data has 54 rows and 15 columns (26 by the end of the report, but I’m not using all the variables). It is what I expected, I entered the information myself into an Excel sheet. Each row of the data represents a book that has been requested. There are a few books that have incomplete data, and I’m still tracking them. I have entered the incomplete information as NA. I have also expanded the data to include more books that come in both regular and large print, since starting the project I questioned whether this would also be interesting. It’s unlikely that I will have more usable data before the due date though.

The variables that I expect to be unique are the titles of the books, except in cases where I have included a large print version to compare its speed with regular print. I don’t consider them duplicates, though they have the same title, because they are listed separately, and have a slightly different audience.

**Preparing the Data for Analysis – Variable Types:**

I do not have any variables that are saved as factor variables. The Date\_Requested variable is a POSIXct variable.

I had issues with Days\_until\_on\_Hold variable. It was showing as a character variable, though it contained integers with some NAs. I made it numeric.

**Preparing the Data for Analysis – Missing Values:**

All missing values are represented with NA.

**Preparing the Data for Analysis – Dealing with Dates:**

I don’t have dates that I want to directly analyze, but I have kept track for time measurement, which I measured in weeks and days.

**Preparing the Data for Analysis – Creating New Variables:**

I made some new variables, Audience\_Refined was created to increase the number of observations for one level, combining children and teen.

Some books have yet to come in, and I have their end date as NA. To still use their speed data, I created a variable to represent their movement from start to three weeks, q\_start\_three\_wks.

q\_daily is the measurement of the speed of a book through the queue per day, and q\_weekly is the weekly speed. I separated out most of the speeds for the different genres to make the code simpler to read.

Is\_popular was created to measure whether something was more popular than the mean of the popular variable, defined as (starting\_position\_in\_queue) / (No.\_of\_Copies). I figured the greater the number, the more popular. I created these variables to use in the investigations of section 4.

Extremes was created to filter the books where they have either a small or large amount of copies.

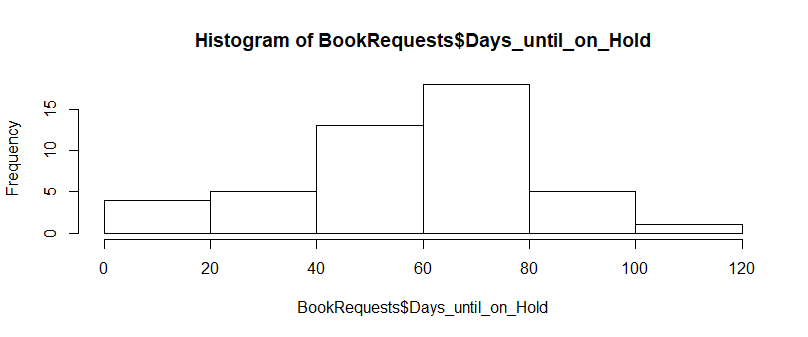
I created a few other variations of variables, some used in this project, and some for later use. I’ll include them in my R code.

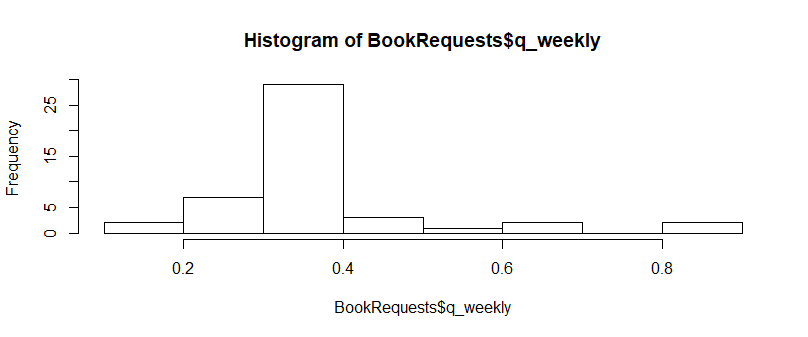
**Exploring Your Data, One Variable at a Time:**

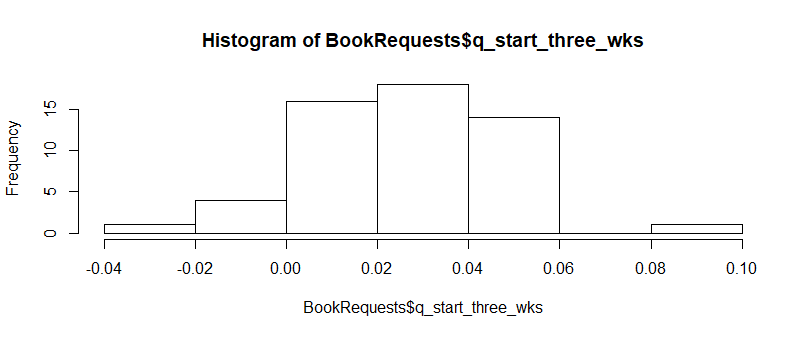
I made histograms of Days\_until\_on\_Hold, q\_weekly, and q\_start\_three\_wks.

For q\_start\_three\_wks I noticed that there were negative values. I had assumed all values would be positive since you would expect after three weeks the book would have moved forward in the queue. I looked at the data, and two non-fiction books that were in high demand, and few copies, were the culprits. It is quite possible that the books went ‘missing’, and there is also the problem of using two databases (public vs staff database, and possible caching issues) that didn’t always match. This problem isn’t a big deal when there are many copies, but when the number of copies is small it can be an issue as in these cases. Either way, I felt these values were probably accurate.

The histogram for q\_weekly shows an outlier greater than 0.8.





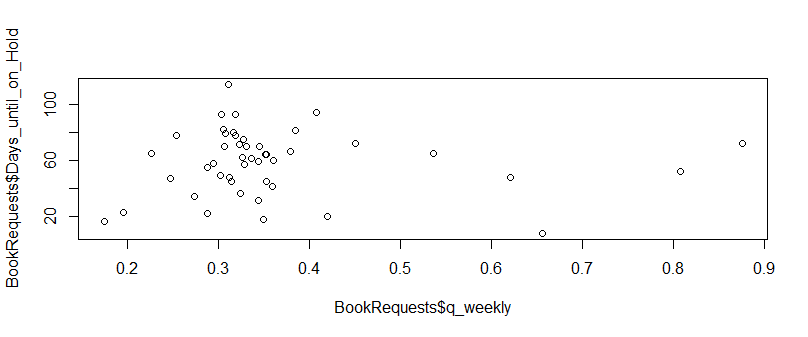


For categorical variables (the three that had interesting information), I made frequency tables:

|  |  |  |
| --- | --- | --- |
| Variable | Levels | N% |
| Audience Refined | Adult fiction | 26 (48.1%) |
|  | Adult non-fiction | 24 (44.4%) |
|  | Non-adult | 4 (7.4%) |
| Regular or Large Print | Large | 3 (5.6%) |
|  | Regular | 51 (94.4%) |
| Fiction or Non-fiction | Fiction | 30 (55.6%) |
|  | Non-fiction | 24 (44.4%) |

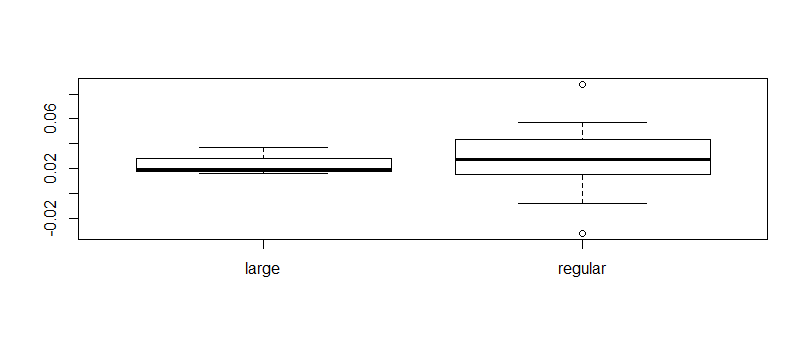
**Visualizing Relationships Between Variables:**

I made a scatterplot of (q\_weekly) vs (Days\_until\_on\_Hold) and (Days\_until\_on\_Hold) vs (No.\_of\_Copies). The first appears to show that most books have a similar speed through the queue. The second appears to show that books with the least number of copies shows the most (quite a lot of) variability, and is something I didn’t think much about. This is something that I think I should consider more.

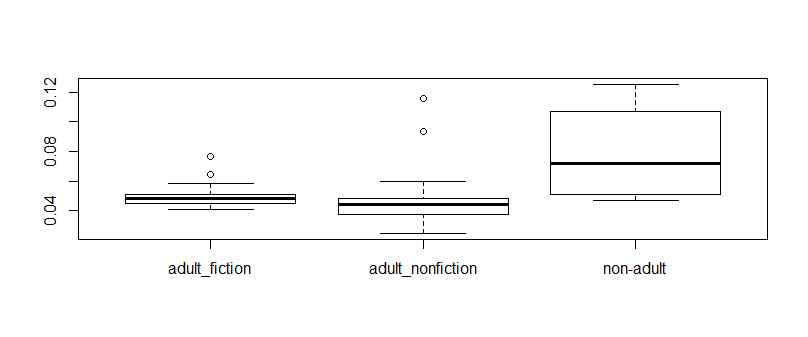




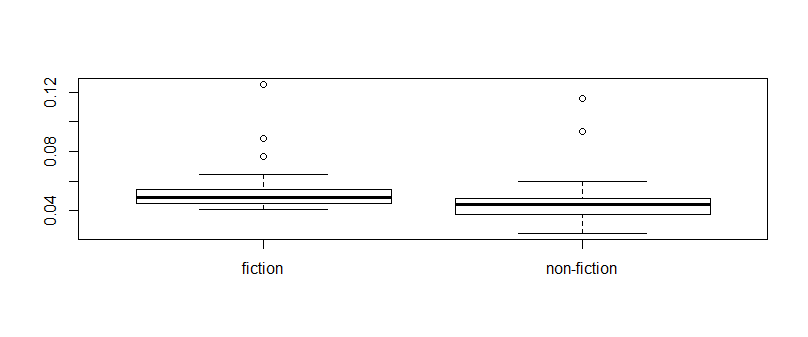
This is a boxplot of the speed through the queue over three weeks for large print and regular print. A better comparison would be between only the books that have a large and regular print.



This is a boxplot of the speed through the queue broken down by adult fiction, adult non-fiction, and non-adult.



This boxplot is the comparison I originally was most interested in working out. The battle between the speed of fiction and non-fiction through the queue.



Section 3

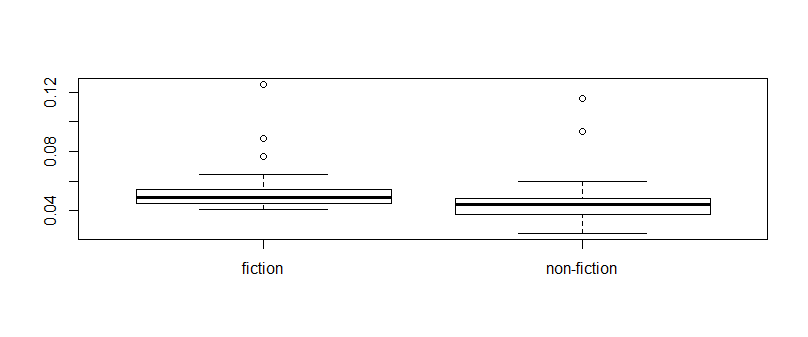
**Plan Your Scientific Questions of Interest:**

* How fast do books move through the queue?
* Is there a difference between fiction and non-fiction, and how fast they move through the queue?
* Is there a difference between the movement of adult non-fiction, adult fiction, and children’s books?

|  |  |  |  |
| --- | --- | --- | --- |
| Analysis Requirements | Name and type of variable 1 | Name and type of variable 2: | Question of interest |
| One numeric variable, one binary variable | Movement (numeric) | Fiction\_or\_Non\_Fiction  (binary) | Is there a difference between fiction and non-fiction, and how fast they move through the queue? |
| One numeric variable, one categorical variable with 3-5 levels | Movement (numeric) | Audience\_Refined (categorical with 3 levels) | Is there a difference between the speed of adult non-fiction, adult fiction, and children’s books through the queue? |

**T-test:**

Here’s the boxplot of the speed of fiction and non-fiction:

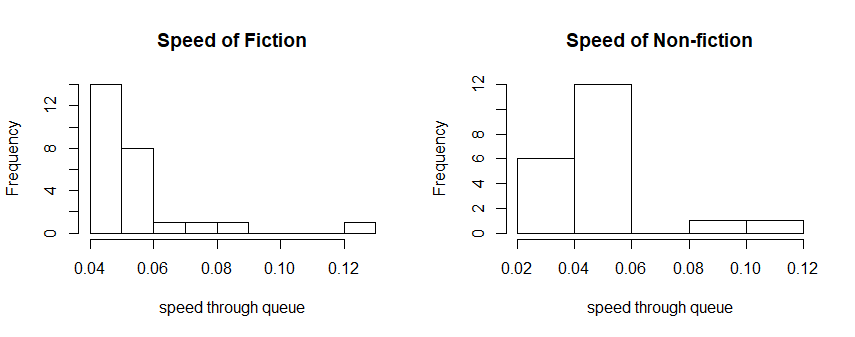


Visually, fiction moves quicker than non-fiction. Both groups look to have outliers, but I know from looking at scatterplots that unusual behavior occurs more often when there are smaller number of copies. Since they both have them, I left it, but it’s something to consider.

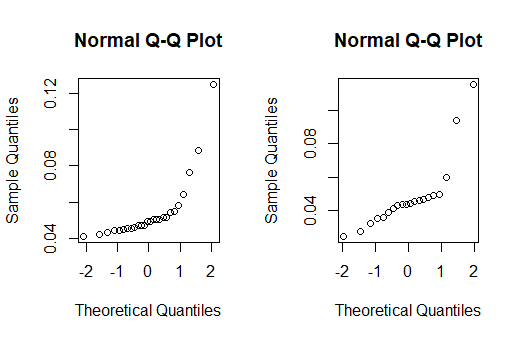
My dataset contains 30 fiction books with an average daily speed of 0.05434, and 24 non-fiction books with an average daily speed of 0.04842.

Check conditions for a t-test:

* Independent observations between and within groups. The observations were taken from a sampling of popular requested books. There are three books that have the same title, but are a different print. They might not be totally independent because sometimes people request both versions to see which is quicker. People also request more than one book at a time, so could also affect independence. I’m assuming independence because I see these as small issues.
* Sample size and skew. One group, non-fiction, is under 30, that data appears more normal than the group has 30 observations. Looked at log transformation because of skew, and it did not improve the skew for fiction.



These are QQ plots, fiction on left and non-fiction on right.



The data is right skewed. I’d rather not remove the outliers though because they occur in both sets, and I feel they represent a real possibility and true values. The number of observations are bordering on being too small, but I will still perform the t-test.

**The null hypothesis is that the speed of fiction books through the queue is equal to the speed of non-fiction. The alternative hypothesis is that they are not equal.**

The point estimate for the difference between the groups is the difference of the means of both groups.

0.05434 - 0.04842 = 0.00592

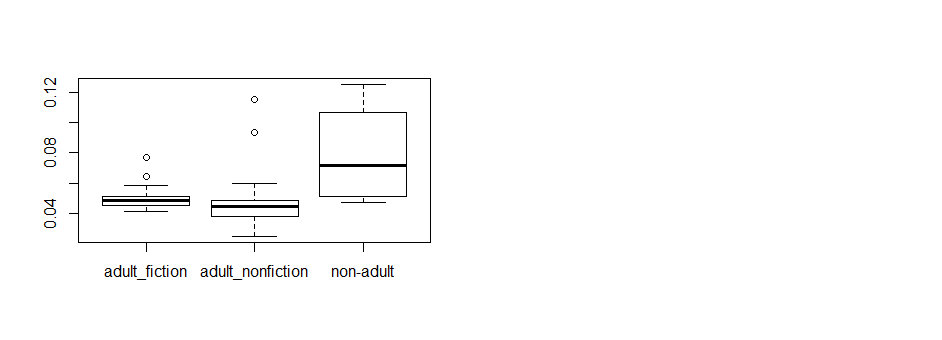
The t-test gives me a p-value of 0.3204.

The 95 percent confidence interval is (-0.05433832, 0.017835181).

The p-value is greater than 0.05, so we do not reject the null. The differences observed were not statistically significant. The null value is within the confidence interval. We are 95% confident that the true mean of the difference between the speed of fiction and non-fiction books lies in the interval between -0.05433832 and 0.017835181.

**ANOVA:**

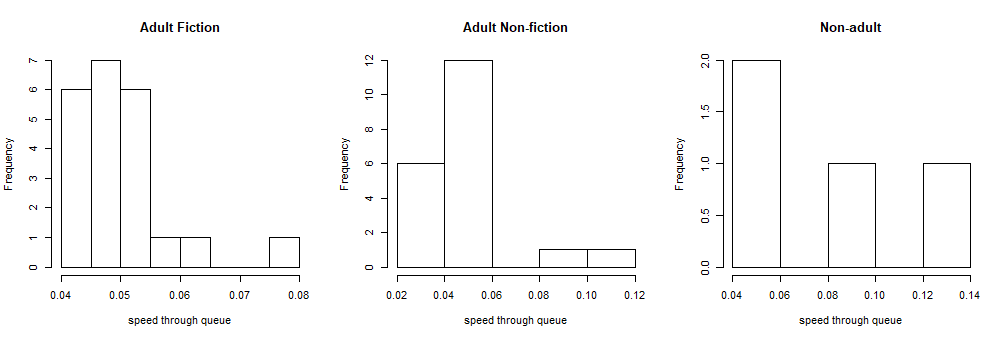
This is a boxplot of the speed of books broken down by intended audience.

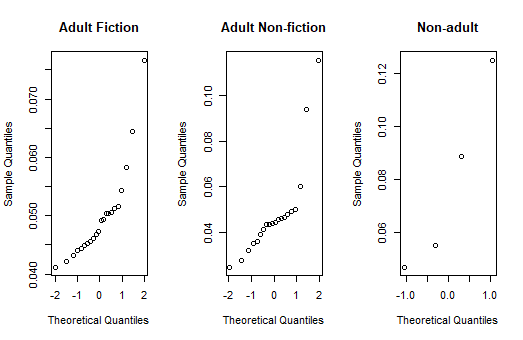


It appears that non-adult moves faster than both adult fiction and non-fiction.

Check conditions for ANOVA:

* Independence between and within groups. I have similar thoughts as I did for the t-test. I will consider them independent.
* Observations are nearly normal within each group. Non-adult only has four observations and does not look normal. Very right skewed and very few observations.
* Equal variability within groups. Looking at the standard deviations, it does not appear to meet this condition. The boxplot spread does not seem as bad.





The QQ-plots show that the data is right skewed. The first two have somewhat of a straight line, then curve on the right. The non-adult plot doesn’t even have that, but it is only four observations.

These are the standard deviations:

|  |  |
| --- | --- |
| Adult Fiction | 0.008049226 |
| Adult Non-fiction | 0.02108637 |
| Non-adult | 0.03563818 |

I will still perform ANOVA, but I feel the conditions are not met, and the results not meaningful.

**The null hypothesis of my ANOVA is that the average speed of the books is the same regardless of audience intended, and genre. The alternative hypothesis is that at least one of the average speeds are different.**

ANOVA table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
| Audience | 2 | 0.0032527 | 0.00162633 | 5.1349 | 0.01 |
| Residuals | 43 | 0.0136189 | 0.00031672 |  |  |

The p-value is less than 0.05. This would usually conclude that at least one average speed is different, and that we should reject the null. I do not think the conditions were met for this test, so I do not find the results meaningful.

The p-value table from the Tukey’s HSD test:

|  |  |
| --- | --- |
| Comparison | P-value |
| adult\_nf vs. adult\_fiction | 0.9622437 |
| non-adult vs. adult\_fic | 0.0121109 |
| non-adult vs. adult\_nf | 0.0086538 |

The non-adult books’ speed differs from both adult fiction and adult non-fiction. There isn’t any evidence for a difference between adult fiction and non-fiction.

My conclusion is that the conditions have not been met for this test, and would need further data to support the results. The t-test also showed there to be no difference between the speeds of fiction and non-fiction books, which suggests at least that part of the test may be accurate.

Section 4 – Plan Your Scientific Questions of Interest

Table 7: Planned scientific questions for Section 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Analysis Requirements | Name and type of variable 1 | Name and type of variable 2 | Question of interest |
| Test of a difference in two proportions | Two binary variables | Extremes (measures whether has number of copies < 20 or >150, binary | Fiction or nonfiction, binary | Is the proportion of books with n < 20 or >150 different whether fiction or non-fiction |
| Chi Sq test of goodness of fit | One categorical variable | Fiction\_or\_Non-Fiction | NA | Does the data support my expectation to have an equal number of fiction and non-fiction books (50/50) |
| Chi Sq test of independence | Two categorical variables | Is\_popular (popular measurement > mean), binary | Audience\_refined, categorical | Are audience intended and popularity independent? |

**Section 4 – Test for a Difference in Two Proportions:**

Table 8: Number and proportion of books with an extreme number of copies (< 20 || >150) by fiction or non-fiction

|  |  |  |
| --- | --- | --- |
|  | Fiction | Non-fiction |
| Extremes | N (column proportion) | N (column proportion) |
| no.\_of\_copies <20 || >150 (T) | 14 (0.47) | 13 (0.54) |
| 20<no.\_of\_copies<150 (F) | 16 (0.53) | 11 (0.46) |
| Total | 30 (1) | 24 (1) |

The conditions for a test of two proportions are independence between groups, independence within groups, and at least 10 expected successes and failures in each group. The books were chosen randomly from requested books, so should be independent between groups and within the groups.

To calculate expected successes, we need the pooled sample proportion of extreme number of copies: (14 + 13) / (30 + 24) = 0.5. The expected number of extremes for the fiction group is (0.5 \*30) = 15. For the non-fiction it would be 12. The expected number of successes and failures in each group is > 10.

**The null hypothesis for the test of two proportions is that the proportion of extreme values will be equal for fiction and non-fiction. The alternative hypothesis is that they are not equal.**

I get a P-value of 0.7842 using prop.test. We do not reject null hypothesis. The sample proportions are likely. The 95% confidence interval is (-0.2300952, 0.3800952). We are 95% confident the difference of proportions of Extremes for fiction and non-fiction is between this interval, it includes 0.

**Section 4 - Test for Goodness of Fit:**

My expectation of the distribution of fiction and non-fiction books is ½ and ½.

Table 9: Frequency table of activity type

|  |  |
| --- | --- |
| Fiction or Non-fiction | N (column proportion) |
| Fiction | 30 (0.56) |
| Non-fiction | 24 (0.44) |

The chi-square test for goodness of fit requires at least 5 observations per cell to be valid, this condition is met. The p-value = 0.4142. This is a large value, and the proportions observed match what I expected.

**Section 4 – Test for Independence Between Two Categorical Variables:**

Table 10: Frequency table of whether an item is popular by its audience. N (column proportion) shown

|  |  |  |  |
| --- | --- | --- | --- |
|  | Adult Fiction | Adult Non-fiction | Non-Adult |
| Is\_popular > mean | 13 | 7 | 3 |
| Is\_popular < mean | 13 | 17 | 1 |
| Total | 26 | 24 | 4 |

The null hypothesis is that the variables are independent. Alternative is that they are not independent.

The non-adult cells are all less than 5, not meeting the chi-square test condition. I ran the test anyway and the p-value = 0.1306. There was a warning from R Studio that the approximation may be incorrect.

I then ran the chi-square test with a different categorical variable (Fiction or Non-fiction), and the p-value = 0.1316, and no warning. The sampling was random, and the variables were categorical. The results are similar, and the only difference between the variables is that the third column (non-adult) was split into the other two groups.

There does not appear to be a significant relationship between these categorical variables. This shows these variables are probably independent, and null hypothesis is not rejected.

**Section 5 – Conclusions:**

I assumed that I would find a relationship between the genre of book, and how quickly it would move through the queue. From my observations, and just intuition, this seemed to make sense. It was not backed up by my data. Any variation was minimal, and not significant. I could gather more data to be sure.

There are some further questions that I’d like to analyze. I would like to compare the speed of the regular and large print versions of the same book. People are often told to request both because large print moves a little faster. Is that true? The problems though, are that there aren’t many books with both versions, and the large print often comes out after the book has already shown to be popular. So, there isn’t as much demand, and they have shorter queues.

How much of an effect does the number of copies have on the wait? If a book goes missing, it’s not a big deal unless you only have five copies. There were some outliers in my data, and I have a feeling this could be part of the problem. Only one way to find out.

There were also database issues. There is a staff catalog and a public catalog. It didn’t occur to me that there could be any differences, and some information was collected with one or the other, or both. I didn’t realize this was a problem until I started inputting data, and noticed that some things were further back in the queue than they should have been. This shouldn’t make a difference overall because I was mainly concerned about the change from start to finish. It did open my eyes to the difference in databases, and caching.

It seems you cannot have too much data. At least in this case. While working on the project, I often kept thinking about why I didn’t think to gather more information. Overall, I feel satisfied with the results and it simplifies my work project ([BookRequests](https://codepen.io/cpooley/full/oExjba/)).

Cyndy Pooley

**R-Code:**

BookRequests <- read\_excel("C:/Users/cynth/Desktop/StThomas/SEIS 631 Data Analysis/BookRequests.xlsx")

View(BookRequests)

dim(BookRequests)

table(BookRequests)

str(BookRequests)

BookRequests$Days\_until\_on\_Hold<-as.numeric(BookRequests$Days\_until\_on\_Hold)

## q\_daily is the measurement of movement of all book through the queue per day

BookRequests$q\_daily <- (BookRequests$Starting\_position\_in\_Queue/BookRequests$No.\_of\_Copies)/BookRequests$Days\_until\_on\_Hold

summary(BookRequests$q\_daily)

BookRequests$q\_weekly <- (7 \* BookRequests$q\_daily) ## weekly movement

summary(BookRequests$q\_weekly)

##movement of adult nonfiction, adult fiction, non-adult

q\_adult\_nf <- BookRequests$q\_daily[BookRequests$Audience\_Refined == 'adult\_nonfiction']

adult\_nf <-(BookRequests$Audience\_Refined == 'adult\_nonfiction')

q\_adult\_fic <- BookRequests$q\_daily[BookRequests$Audience\_Refined == 'adult\_fiction']

q\_non\_adult <- BookRequests$q\_daily[BookRequests$Audience\_Refined == 'non-adult']

summary(q\_adult\_nf)

summary(q\_adult\_fic)

summary(q\_non\_adult)

##movement of large print, regular print

##large print only has three values, one of which is NA since it doesn't have a

##value for Days\_until\_on\_hold

##create variables for the change from start to three weeks, so I can have third value

q\_reg\_print <- BookRequests$q\_daily[BookRequests$Regular\_or\_Large\_Print == 'regular']

q\_large\_print <- BookRequests$q\_daily[BookRequests$Regular\_or\_Large\_Print == 'large']

summary(q\_reg\_print)

summary(q\_large\_print)

BookRequests$start\_three\_wks <- BookRequests$Starting\_position\_in\_Queue - BookRequests$Position\_After\_3\_Weeks

BookRequests$q\_start\_three\_wks <-(BookRequests$start\_three\_wks/BookRequests$No.\_of\_Copies)/ 21

q\_start\_threewks\_large\_print <- BookRequests$q\_start\_three\_wks[BookRequests$Regular\_or\_Large\_Print == 'large']

q\_start\_threewks\_reg\_print <- BookRequests$q\_start\_three\_wks[BookRequests$Regular\_or\_Large\_Print == 'regular']

##histograms

hist(BookRequests$Days\_until\_on\_Hold)

hist(BookRequests$q\_weekly)

hist(BookRequests$q\_start\_three\_wks)

##The second histogram has a negative value, meaning some values went backwards, checked data and two items had moved back, these are possibly correct

summary(BookRequests$q\_start\_three\_wks)

min(BookRequests$q\_start\_three\_wks)

##frequency tables

table(BookRequests$Regular\_or\_Large\_Print)

table(BookRequests$Audience\_Refined)

table(BookRequests$`Fiction\_or\_Non-Fiction`)

##proportions

prop.table(table(BookRequests$Regular\_or\_Large\_Print))

prop.table(table(BookRequests$Audience\_Refined))

prop.table(table(BookRequests$`Fiction\_or\_Non-Fiction`))

##scatterplots

plot(BookRequests$Starting\_position\_in\_Queue, BookRequests$Position\_After\_3\_Weeks)

plot(BookRequests$No.\_of\_Copies, BookRequests$Days\_until\_on\_Hold)

##The ones with the shortest and longest days on hold are the ones with the fewest no of copies

plot(BookRequests$q\_weekly, BookRequests$q\_start\_three\_wks)

plot(BookRequests$q\_weekly, BookRequests$No.\_of\_Copies)

plot(BookRequests$Starting\_position\_in\_Queue, BookRequests$No.\_of\_Copies)

##I like this next version better

plot(BookRequests$Days\_until\_on\_Hold, BookRequests$No.\_of\_Copies)

plot(BookRequests$Days\_until\_on\_Hold, BookRequests$q\_weekly)

##This next one is better

plot( BookRequests$q\_weekly, BookRequests$Days\_until\_on\_Hold)

##boxplots

boxplot(BookRequests$q\_start\_three\_wks ~ BookRequests$Regular\_or\_Large\_Print)

boxplot(BookRequests$q\_daily ~ BookRequests$Audience\_Refined)

boxplot(BookRequests$q\_daily ~ BookRequests$`Fiction\_or\_Non-Fiction`)

##finding mean for fiction, non-fiction speed dataset

q\_fiction <-(BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction'])

summary(q\_fiction)

table(BookRequests$`Fiction\_or\_Non-Fiction`)

q\_nonfiction <-(BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'non-fiction'])

summary(q\_nonfiction)

hist(BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction')

##side by side histograms

par(mfrow = c(1,2))

hist(BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction'], main = 'Speed of Fiction', xlab = 'speed through queue')

hist(BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'non-fiction'], main = 'Speed of Non-fiction', xlab = 'speed through queue')

##qqnorm

##data is skewed, trying log transformation

par(mfrow = c(1,1))

hist(log(BookRequests$q\_daily))

par(mfrow = c(1,2))

hist(log(BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction']))

hist(log(BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'non-fiction']))

##Fiction still looks very skewed, but non-fiction looks more normal

par(mfrow = c(1,2))

qqnorm((BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction']))

qqnorm((BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'non-fiction']))

##t-test

t.test(x =BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction'], y = BookRequests$q\_daily[BookRequests$`Fiction\_or\_Non-Fiction`== 'non-fiction'], alternative = 'two.sided')

##start of ANOVA

boxplot(BookRequests$q\_daily ~ BookRequests$Audience\_Refined)

##plots to assess normality, created variables earlier

par(mfrow = c(1,3))

hist(q\_adult\_fic, main = 'Adult Fiction', xlab = 'speed through queue')

hist(q\_adult\_nf, main = 'Adult Non-fiction', xlab = 'speed through queue')

hist(q\_non\_adult, main = 'Non-adult', xlab = 'speed through queue')

##qqplot

qqnorm(q\_adult\_fic, main = 'Adult Fiction')

qqnorm(q\_adult\_nf, main = 'Adult Non-fiction')

qqnorm(q\_non\_adult, main = 'Non-adult')

##standard deviations, remove NAs

sd(q\_adult\_fic, na.rm = T)

sd(q\_adult\_nf, na.rm = T)

sd(q\_non\_adult)

##ANOVA

fit <- lm(BookRequests$q\_daily ~ BookRequests$Audience\_Refined)

anova(fit)

TukeyHSD(aov(fit))

##create new variables

##slow\_q binary variable, less than mean speed(0.05176), true is slow

BookRequests$slow\_q <- (BookRequests$q\_daily < 0.05176)

summary(BookRequests$q\_daily)

table(BookRequests$slow\_q)

##is\_popular variable, binary, estimate popularity by starting position/number of copies, the greater the number the more popular

BookRequests$popular <- (BookRequests$Starting\_position\_in\_Queue)/(BookRequests$No.\_of\_Copies)

summary(BookRequests$popular)

##mean value for popular is 3.340, so > 3.340 is true

BookRequests$is\_popular <- (BookRequests$popular > 3.340)

summary(BookRequests$is\_popular)

summary(BookRequests$is\_popular[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction'])

summary(BookRequests$is\_popular[BookRequests$`Fiction\_or\_Non-Fiction`!= 'fiction'])

##new variable, low\_copies, T or F, no.\_of\_copies<=30

BookRequests$low\_copies <- BookRequests$No.\_of\_Copies <= 80

summary(BookRequests$low\_copies[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction'])

summary(BookRequests$low\_copies[BookRequests$`Fiction\_or\_Non-Fiction`== 'non-fiction'])

##another variable

BookRequests$Extremes <- BookRequests$No.\_of\_Copies < 20 | BookRequests$No.\_of\_Copies >150

summary(BookRequests$Extremes[BookRequests$`Fiction\_or\_Non-Fiction`== 'fiction'])

summary(BookRequests$Extremes[BookRequests$`Fiction\_or\_Non-Fiction`!= 'fiction'])

##Test of two proportions

table(BookRequests$Extremes, BookRequests$`Fiction\_or\_Non-Fiction`)

prop.table(table(BookRequests$Extremes, BookRequests$`Fiction\_or\_Non-Fiction`), 2)

##prop.test(table(BookRequests$Extremes, BookRequests$`Fiction\_or\_Non-Fiction`))

prop.test(table(BookRequests$`Fiction\_or\_Non-Fiction`, BookRequests$Extremes))

chisq.test(table(BookRequests$Extremes, BookRequests$`Fiction\_or\_Non-Fiction`))

##Chi qs test of goodness of fit

table(BookRequests$`Fiction\_or\_Non-Fiction`)

prop.table(table(BookRequests$`Fiction\_or\_Non-Fiction`))

chisq.test(table(BookRequests$`Fiction\_or\_Non-Fiction`), p = c(1/2, 1/2))

##Chi sq test for independence

table(BookRequests$is\_popular, BookRequests$Audience\_Refined)

table(BookRequests$is\_popular, BookRequests$`Fiction\_or\_Non-Fiction`)

chisq.test(table(BookRequests$is\_popular, BookRequests$Audience\_Refined))

chisq.test(table(BookRequests$is\_popular, BookRequests$`Fiction\_or\_Non-Fiction`))