# Research Skills, Experimental Methods

Robert Hoff

November 24, 2011

## 1 Analysis

Considering the results visually. Figure 1 shows a comparison of all the 35 samples for each of the search engines. These each look faster in turn, so a difference in performance seems likely. Figure 2 shows a similar comparison for each of the days, we can note some slight variation, in particular Thursday and Friday looks slower and faster respectively. We want to know if there might be some deviations or systematic errors in the underlying data. Figure 3 shows the performance of the three engines for each of the days, we will expect that each of the days be consistent with what the complete data suggests. Similarly, figure 4 shows the samples for the days for each of the engines, if it turns out that Friday is significantly faster than the other days, this should hold true for each of the separate engines also.

A consideration is to look for consistent relationships between the engines. Figure 5 compares the mean of alpha and nu for each of the days. This result almost certainly looks inconclusive. We would also like to know if the sample set for a given engine vary normally for all of the days. Figure 6 does resemble a normal distribution for the set. To test this we produce a QQ plot shown in Figure 7.

## 2 Report

### 2.1 Design

The samples were collected by a computer program that issues the searches using a query string by http request, it measures the time precisely by the time it takes to receive the http response. To cancel out the network latencies between each of the three search engines, we issue in parallel a ping request which is subtracted from the measurement. We constructed each of the 105 queries uniquely from random dictionary words of between 2 and 5 word length. With this approach there will be no biases between any of the query sets. The samples were programmed to be taken at equally spaced intervals throughout each day.

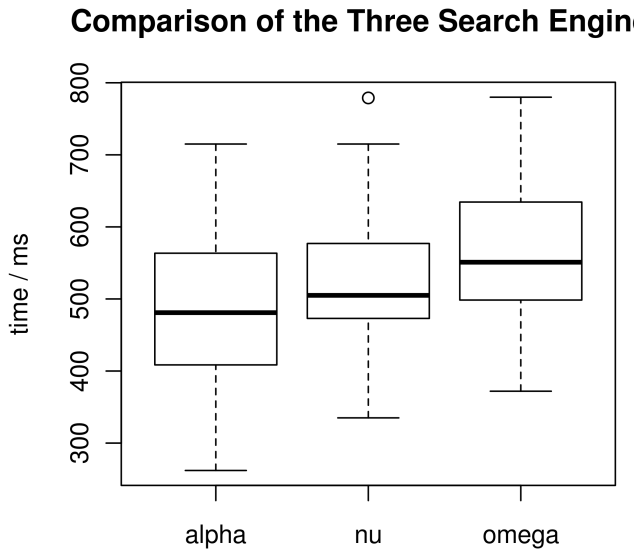


Figure 1: Comparison of 35 samples for each of the search engines

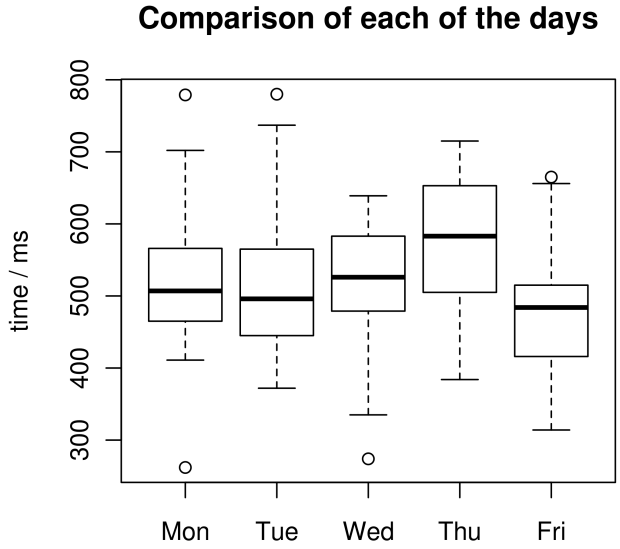


Figure 2: Comparison of 21 samples for each of the five days

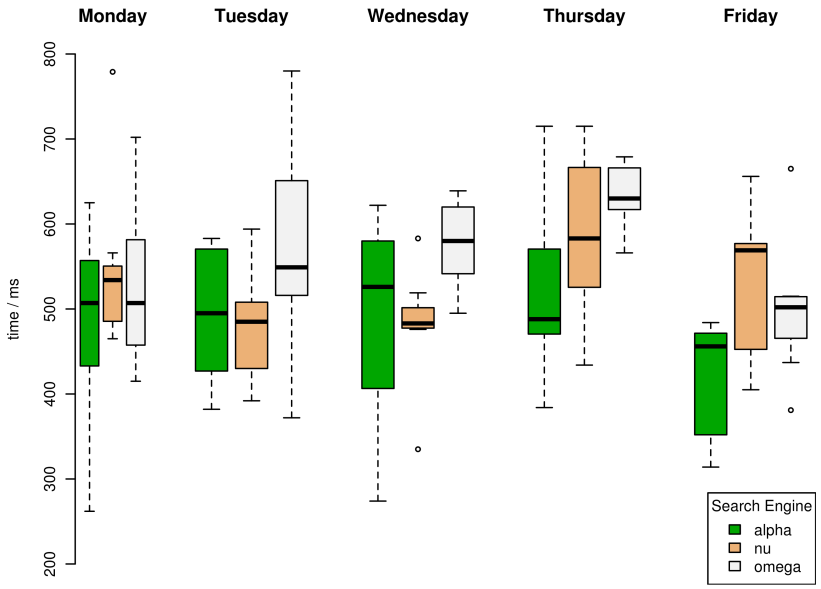


Figure 3: Comparison of the search engines for each of the days

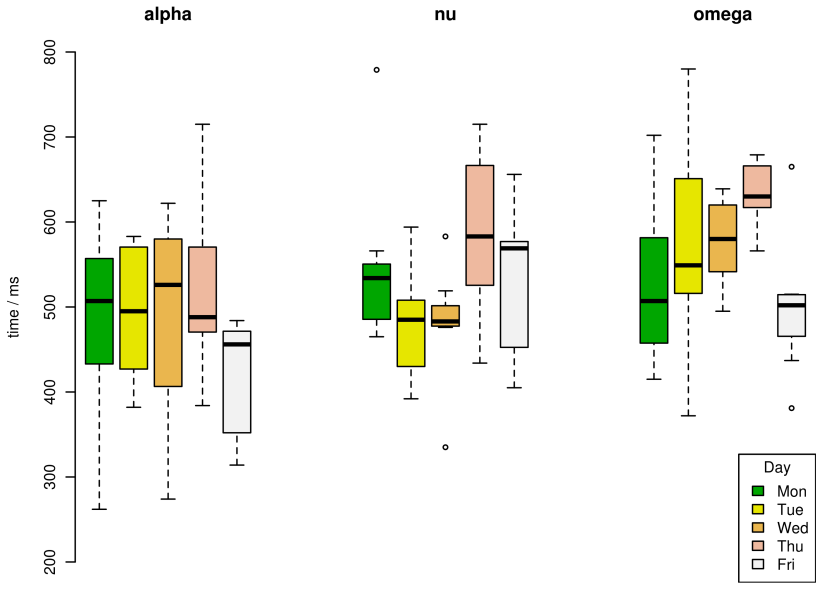


Figure 4: Comparison of the days for each of the search engines

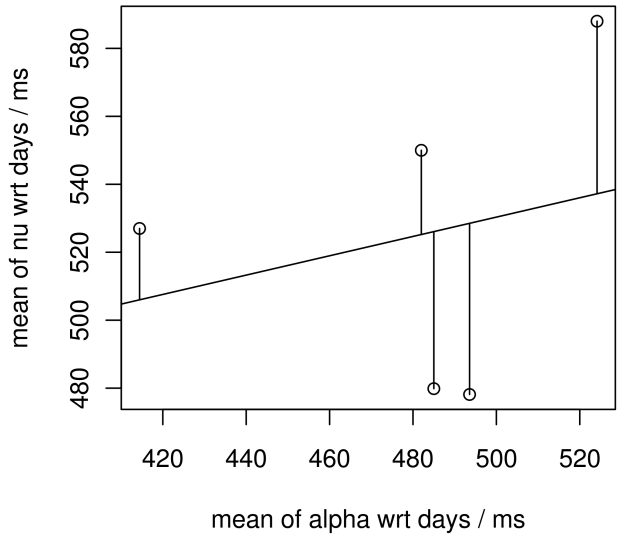


Figure 5: Inspecting possible relationship in load between two search engines for each of the five days

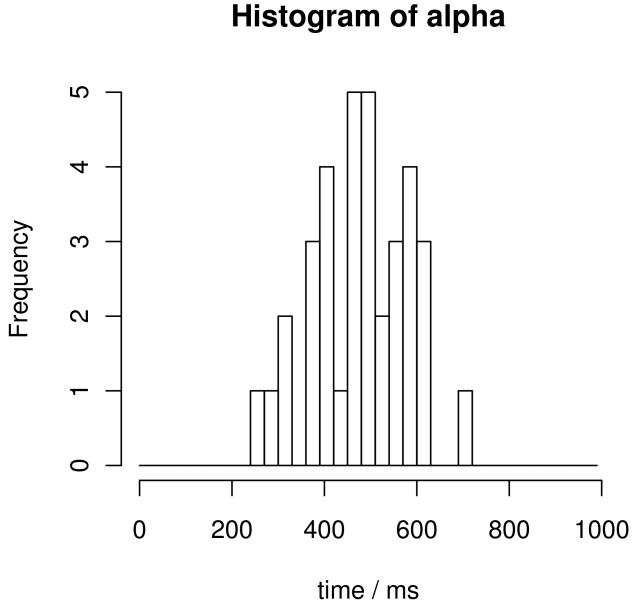


Figure 6: Histogram of the 35 samples by alpha, at 30ms intervals

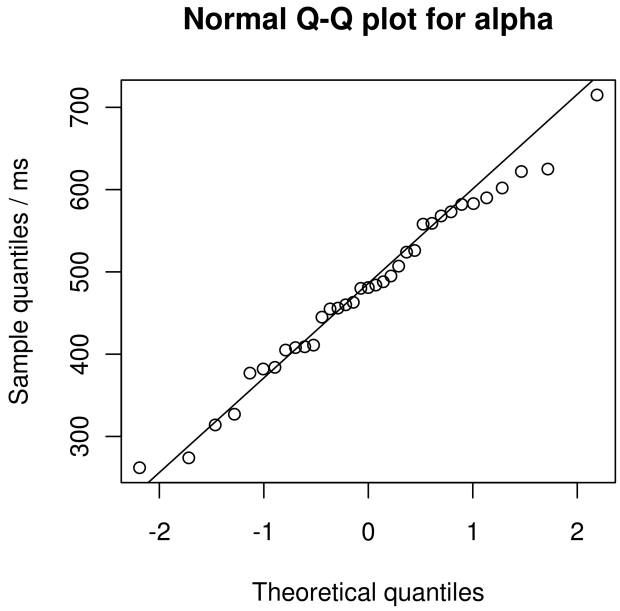


Figure 7: The horizontal axis shows number of standard deviations from the mean

## 2.2 Results

We use the t tests for pairwise relationships. An $H_1$ hypothesis is formulated for a selection of relationships shown in table 1. We can conclude at the 0.05 level that omega is slower than nu which is slower than alpha. This also means that omega must be slower than alpha (to an even higher confidence level). We see from the table that there is a significant difference between the results on Thursday and Friday, but differences are not significant for all the days. To verify that none of the individual results conflict with the complete sample tests between engines on particular days were made, and between days on particular engines. It's not possible to find any significant relationship that conflicts with the overall results.

Anova analysis was used on each of the groups to determine overall significance. Results are shown in table 2. An $H_1$ hypothesis that there is significant difference in each of the groups can be accepted. This conclusion is much stronger with the search engines with a larger F value and $p < 0.01$.

## 2.3 Discussion

When we made the analysis on both the search engines and the days, we used the complete set of samples. For instance when analysing the search engines we were using samples from across all five days. This should be ok if we can verify that loads from different days are affecting all the three search engines in the same way, which is probably a fair assumption. If, alternatively, we had chosen to limit our sample sizes to a short time-frame, other type of biases could have been present. One of the

| $H_1$ | | Confidence |
|---|---|---|
| $\mu(nu) > \mu(alpha)$ | accept | $p < 0.05$ |
| $\mu(omega) > \mu(nu)$ | accept | $p < 0.05$ |
| $\mu(Thu) > \mu(Fri)$ | accept | $p < 0.01$ |
| $\mu(Mon) > \mu(Tue)$ | reject | $p = 0.450$ |
| $\mu(nu.thu) > \mu(omega.thu)$ | reject | $p = 0.312$ |

Table 1: We can conclude that the search engines are significantly different in performance, and Friday is significantly slower than Thursday

| Group | F-value | $p$ |
|---|---|---|
| Search Engines | 6.2247 | 0.002814 |
| Days | 2.8032 | 0.02974 |

Table 2: ANOVA analysis

engines may have been experiencing a technical problem, affecting it adversely for a short duration. We can use a similar reasoning when comparing the days, as long as the engines remain the same for the duration of the experiment, then any bias should cancel out.

None of the q1 to q7 were biased consistently relative to each other so we assumed that they must all have been achieved randomly. Also assumed that they were not a mean over many samples because if the case they would have been more close in value.

There is no reason why we wouldn't want to take more samples in this kind of experiment, since the cost of each sample is zero. If the sample set had been larger we would have been able to investigate more relationship, and make more firm conclusions. This would include looking for consistent performance differences between engines, and being able to obtain a better measure of the actual differences. We would also be able to verify if a normal distribution holds for the sample set for a given search engine or day.

# 3 Data Logs

The following can be downloaded from http://www.cl.cam.ac.uk/ rjh209/

```
data <- c(
# alpha
625, 524, 455, 507, 590, 411, 262,          # monday
382, 445, 495, 559, 409, 583, 582,          # tuesday
602, 405, 408, 558, 526, 274, 622,          # wednesday
384, 573, 715, 568, 481, 488, 460,          # thursday
456, 463, 480, 377, 314, 327, 484,          # friday

# nu
779, 465, 566, 534, 535, 470, 501,
```

```
392, 485, 406, 520, 454, 496, 594,
483, 583, 335, 484, 476, 519, 479,
546, 505, 583, 715, 686, 434, 647,
656, 574, 580, 416, 569, 405, 489,

# omega
507, 702, 433, 612, 551, 482, 415,
372, 492, 565, 737, 540, 780, 549,
536, 495, 639, 580, 601, 547, 639,
630, 653, 621, 679, 566, 679, 613,
515, 665, 437, 502, 514, 494, 381)

alpha <- data[1:35]
nu    <- data[36:70]
omega <- data[71:105]

# holding a day constant (monday)
alpha_m <- data[1:7]
nu_m    <- data[36:42]
omega_m <- data[71:77]
engines_m = gl(3,7, labels=c("am", "nm", "om"))
# plot(engines_m, c(alpha_m, nu_m, omega_m))
# t.test(nu_m, alpha_m, alternative="greater")


mon <- c(data[1:7], data[36:42], data[71:77])
tue <- c(data[8:14], data[43:49], data[78:84])
wed <- c(data[15:21], data[50:56], data[85:91])
thu <- c(data[22:28], data[57:63], data[92:98])
fri <- c(data[29:35], data[64:70], data[99:105])


engines <- gl(3, 35, labels=c("alpha","nu","omega"))
engines_sd <- gl(3, 7, labels=c("a","n","o"))
days <- gl(5, 7, 105, labels=c("Mon","Tue","Wed","Thu","Fri"))
days_se <- gl(5, 7, 35, labels=c("Mon","Tue","Wed","Thu","Fri"))
queries <- gl(7, 1, 105, labels=c("q1","q2","q3","q4","q5","q6","q7"))


# box-plots of alpha(M-F), nu(M-F) and omega(M-F)
plot(engines:days, data, ylim=c(200,800))


# box-plots of monday(a-o), tuesday(a-o), etc
```

```
plot(days:engines, data)

# gives a box-plot for each of the search engines
plot(engines, data, ylab="time / ms")
title("Comparison of the three search engines")

# box-plot for each of the five days
plot(days, data, ylab="time / ms")
title("Comparison of each of the days")


# the mean of the engines across days
agg <- aggregate(data, list(M=days, SE=engines), mean)
alpha_mean <- agg$x[agg$SE=="alpha"]
nu_mean <- agg$x[agg$SE=="nu"]
omega_mean <- agg$x[agg$SE=="omega"]


# inspecting load between alpha and nu for each of the five days
plot(alpha_mean, nu_mean, xlab="mean of alpha wrt days / ms",
                  ylab="mean of nu wrt days / ms")
m<-lm(nu_mean~alpha_mean)
abline(m)
segments(alpha_mean,fitted(m),alpha_mean,nu_mean)
summary(m)

# these results are certainly inconclusive
hist(resid(m),prob=T)
lines(density(resid(m)))


# by visual inspection, the total number of samples may be
# normally distributed
scale <- seq(0, 1000, 30)
hist(alpha, breaks=scale, main="Histogram of alpha", xlab="time / ms")


# investige the possible normal relationship with a Q-Q plot
qqnorm(alpha, main="Normal Q-Q plot for alpha",
                  ylab="Sample quantiles / ms", xlab="Theoretical quantiles")
qqline(alpha)


# the t.tests gives confidence that there is different performance
```

```
# of the search engines
t.test(nu, alpha, alternative="greater")
t.test(omega, nu, alternative="greater")
t.test(thu, fri, alternative="greater")
t.test(mon, tue, alternative="greater")


# finally anova tests make predictions between the means of the groups
summary(aov(data~engines))
summary(aov(data~days))
```