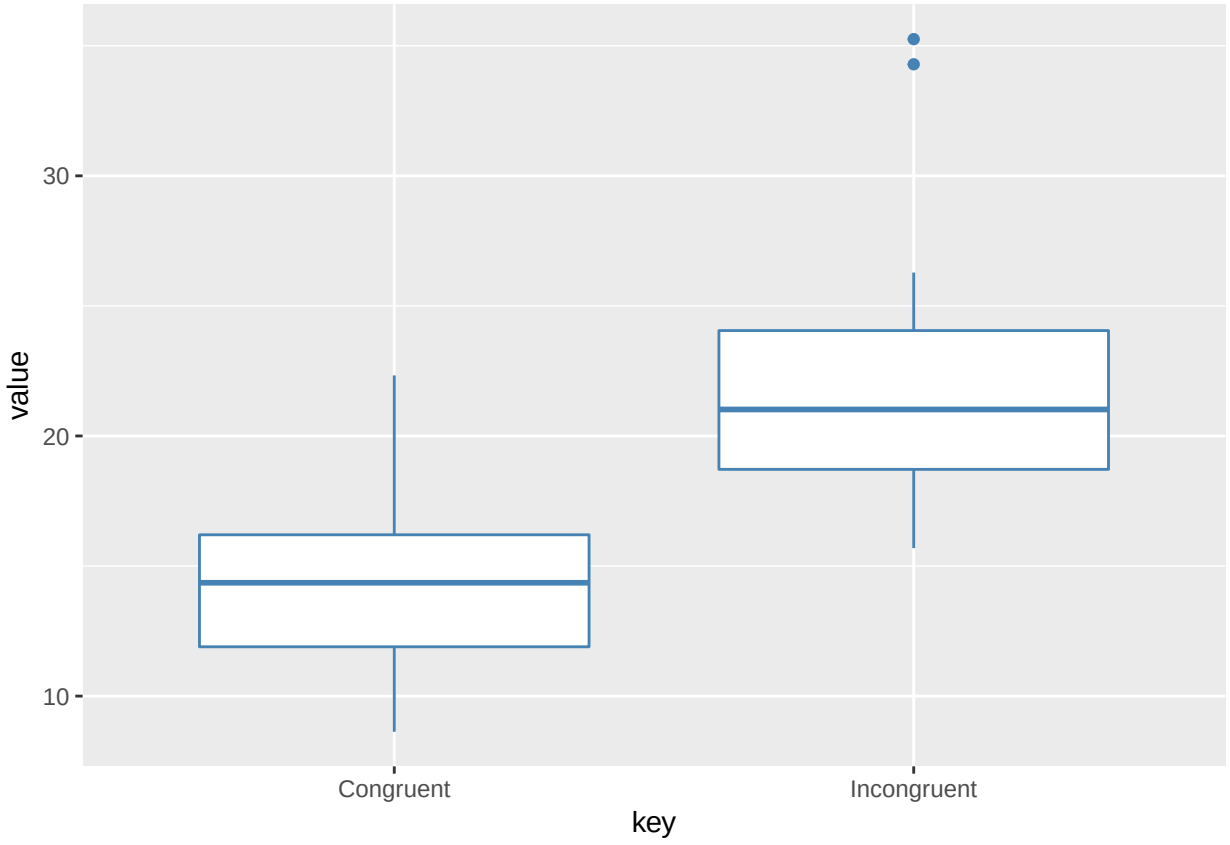
Backround

In a Stroop task, participants are presented with a list of words, with each word displayed in a color of ink. The participant's task is to say out loud the color of the ink in which the word is printed. The task has two conditions: a congruent words condition, and an incongruent words condition. In the congruent words condition, the words being displayed are color words whose names match the colors in which they are printed: for example RED, BLUE. In the incongruent words condition, the words displayed are color words whose names do not match the colors in which they are printed: for example PURPLE, ORANGE. In each case, we measure the time it takes to name the ink colors in equally-sized lists.

Explorative Analysis

First, let's take a look at the data which shows for 24 people the time it took to name the ink colors in equally-sized lists. With the mean being 14.05 for congruent lists and 22.02 for incongruent ones figure 1 implies that naming the font color of a printed word is an easier and quicker task if word meaning and font color are not incongruent.



Next, lets check if we can confirm this picture using simple methods of inferential statistics.

Inferntial Statistics

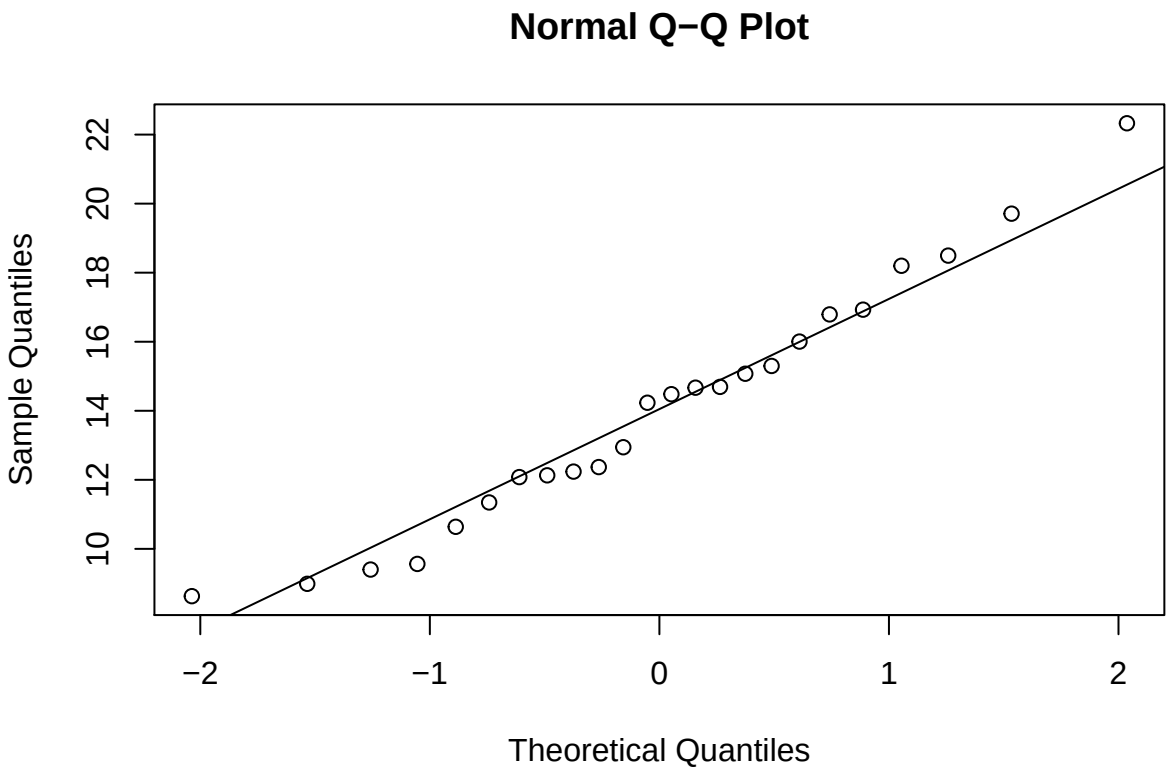
First, let's discuss what we actually want to do and which statistical test is appropriate. What we want to do here is testing the Stroop effect. In order to do so we compare the mean of two dependent samples, that is, paired measurements for one set of items. Our dependent variable is time it takes to name the ink colors and our independent variable is type of list (congruent/incongruent). An appropriate and general formulation of hypotheses for this task may therefore look like the following:

- H0: There is no difference between the two population means of time taken to name the ink colors
- H1: There is a difference between the two population means of time taken to name the ink colors

To test the above stated a paired t-test is appropriate. This test, is used to determine whether the mean of a dependent variable (e.g. reaction time) is the same in two related groups (e.g., two groups of participants that are measured at two different "time points" or who undergo two different "conditions"). In order for the test to be valid the following 4 assumptions must be met (see https://statistics.laerd.com/stata-tutorials/paired-t-test-using-stata.php).

1. The dependent variable should be measured at the interval or ratio level (i.e., they are continuous).
2. The independent variable should consist of two categorical, "related groups" or "matched pairs".
3. There should be no significant outliers in the differences between the two related groups.
4. The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed.

The study design and choice of variables clearly match assumptions 1 and 2. Regarding outliers the boxplots showed that in the incongruent group there are two people who were extraordinarily slow, though, arguably the values are still ok. Assumptions 4 is to my mind best investigated using a simple Q-Q plot.



From the plot above, we see that the distribution is not significantly different from the normal distribution. In other words, we can assume normality. Note that, the test is quite robust to violations of normality, meaning that the assumption can be a little violated and still provide valid results.

Finally let's do the test which is a peace of cake with R.

```
##
##  Paired t-test
##
## data:  strato$value by strato$key
## t = -8.0207, df = 23, p-value = 4.103e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.019028  -5.910555
## sample estimates:
## mean of the differences
##               -7.964792
```

The output provides useful descriptive statistics for the two groups that we compared as well as actual results from the paired t-test. Looking at the mean, we can see that there is a mean difference between the two trials of 7.964 seconds As the p-value is less than 0.05 (meaning that the difference between our to means is unlikely to have occurred because the sample happened to be atypical), it can be concluded that there is a statistically significant difference between our two means. More precisely, assuming that there is no Stroop effect, we would obtain the observed difference or more in only 4.103e-08 percent of studies due to random sampling error. Thus, I reject the null hypothesis.