Sampling based approaches

T. Florian Jaeger

2/26/2020

Sampling-based approaches

We can distinguish between two approaches to the statistical inferences scientists try to draw about the world from the samples they obtain in their experiments. For example, how do we quantify the mean difference in reaction time between the high vs. low frequency condition of our example scenario? One approach is analytic, using mathematical knowledge about the distributions of our variables. For example, the mean of the difference between two normally distributed (non-correlated) variables is the difference between the means of the two distributions; the standard deviation of the difference is the square root of the sum of the variances of the two distributions. So if we assume that reaction times (RTs) are normally distributed, and if high frequency objects would lead to an RT distribution with mean 250 and standard deviation 21, and low frequency objects would lead to an RT distribution with mean 293 and standard deviation 32, then the difference between the two RT distributions would be distributed with mean 250 - 293 = `-43` and standard deviation sqrt(21^2 + 32^2) = `38.2753184`.

The second approach is the **sampling-based approach**. For this approach, we repeatedly draw samples from the two distributions and calculate the difference between each pair of samples (this is not exactly the same use of the word sample as in the previous section where we referred to samples of observations we collect in experiments; but the two uses are related). These differences then form a distribution, and we can get the mean and standard deviation of that distribution. More generally, we can obtain any quantity from the distribution—even more complex ones—by repeatedly sampling from it. As we'll show below, this lets us, for example, construct confidence intervals, or it lets us assess how likely we would be by chance to see the differences we observe in our sample (and thus how informative these differences are). The sampling-based approach is particularly powerful when we do not the analytic solution (or are not sure about it).

To make this more concrete, let's go through an example.

Building intuition about sampling: an example

As an example, let's simulate *one* instance of the hypothetical experiment described above, with a total of 10,000 pairs of observations—one each for each pair of a low and high frequency object. For now, let's assume that we're drawing these 10,000 from the same (very patient) subject.

In the code chunk below you can see what mean and standard deviation we are sampling from. In this case, we *know* the parameters of the simulated (population) data. For our experiments, this is, of course, not the case. Rather we want to *infer* the true parameters. We get to that later. Here our first goal is to build some intuition about sampling.

```
# Let's draw 10000 samples of RTs for both the high and the low frequency distribution.
# We again assume normality and the same means and SDs we assume above.
RT.high = rnorm(n = 10000, mean = 250, sd = 21)
RT.low = rnorm(n = 10000, mean = 293, sd = 32)
```

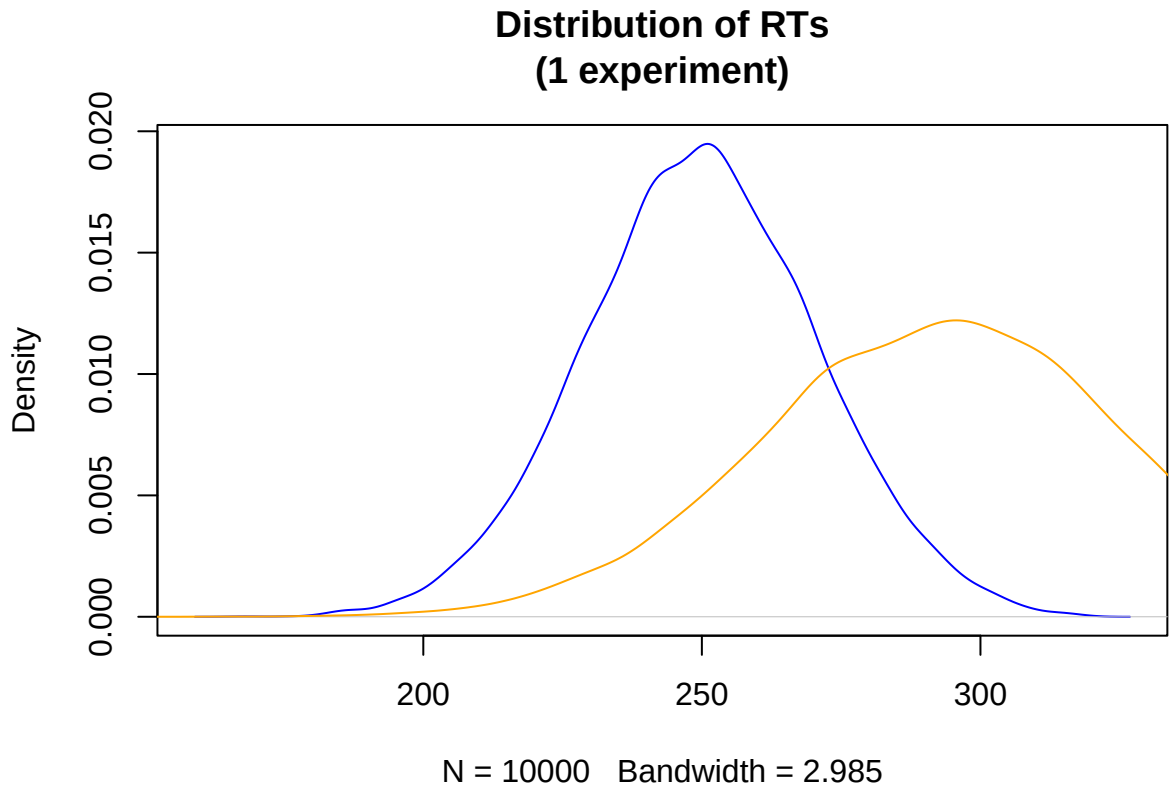
```
# Let's pairwise subtract the RTs
RT.difference = RT.high - RT.low
```

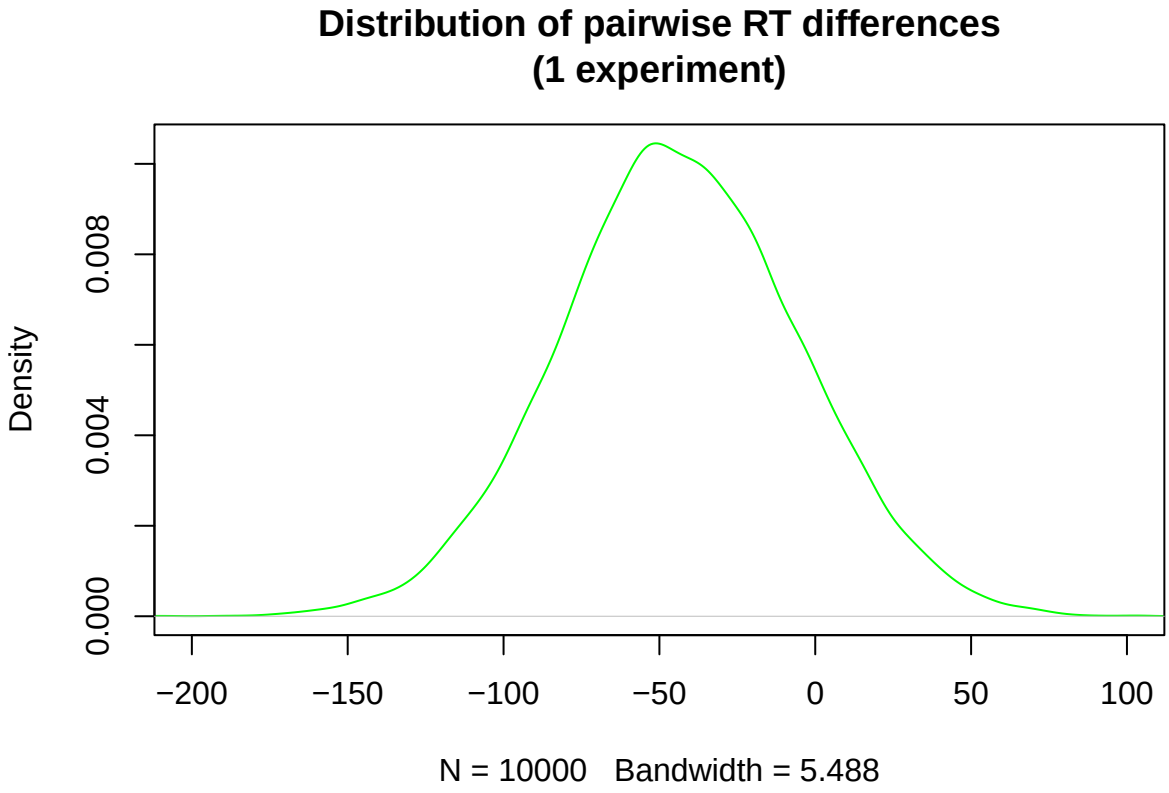
Let's look at the first 10 of these RT differences

```
head(paste(RT.high, RT.low, RT.difference, sep = "; "), 10)
```

```
##  [1] "234.294232639514; 328.593234720487; -94.2990020809736"
##  [2] "266.657295209366; 292.002289749858; -25.3449945404925"
##  [3] "306.565878603572; 272.456690301619; 34.109188301953"
##  [4] "246.214079949705; 256.52365488567; -10.3095749359651"
##  [5] "259.509513577399; 279.06020003305; -19.5506864556512"
##  [6] "260.526491062301; 298.613350177057; -38.0868591147557"
##  [7] "215.396807093741; 190.328119253902; 25.0686878398385"
##  [8] "308.517503390934; 297.681287541609; 10.8362158493245"
##  [9] "280.282706553621; 227.682827258308; 52.5998792953138"
## [10] "262.450259233682; 270.946192448016; -8.49593321433412"
```

And let's plot the distributions of the RTs for low and high frequency objects, as well as the distribution of their differences (taken within each pair of objects):





Since this is a distribution, we can obtain the mean and standard deviation of that distribution:

```
mean(RT.difference)
```

```
## [1] -43.60291
```

```
sd(RT.difference)
```

```
## [1] 38.52763
```

Notice how similar these numbers are to those we obtained analytically above. The advantage of the sampling based approach, however, is that we can also apply it when the analytic solution is not readily available! That makes it a powerful tool for us to understand our data by sampling or **simulating** the quantities we are trying to understand.

Repeated sampling: continuing our example

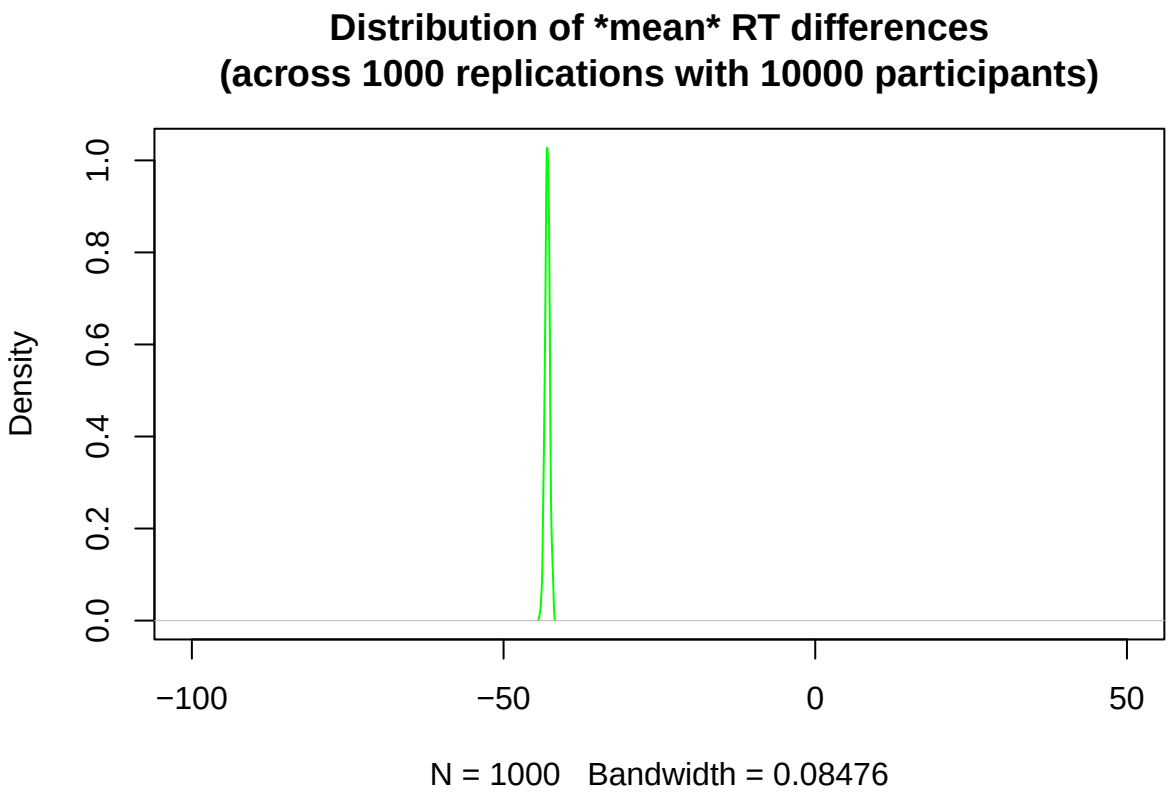
What we've done so far is to take one sample—one simulation of a thought experiment. Let's do that many times, so that we can see what *distribution of the mean* RT differences we would expect:

```
# A function that makes one simulation with 10000 samples and returns the 10000
# differences in RTs
my_sample = function(n = 10000) {
  RT.high = rnorm(n = n, mean = 250, sd = 21)
  RT.low = rnorm(n = n, mean = 293, sd = 32)

  RT.difference = RT.high - RT.low
}
```

```r
# Let's take the mean of 10000 RT differences, and let's repeat that 1000 times
means.RT.difference = replicate(1000, mean(my_sample()))

head(means.RT.difference, 20)
```

```
##  [1] -42.36174 -42.54281 -42.53833 -42.81841 -43.19431 -42.38652 -43.22289
##  [8] -42.46736 -42.66477 -42.98617 -42.85368 -43.16889 -42.43231 -42.59353
## [15] -43.19038 -43.14337 -43.06249 -42.57276 -42.73381 -43.67853
```

```r
# The distribution of the *mean* differences across repeated samples
# (repeated draws of our thought experiments)
plot(density(means.RT.difference), col = "green", xlim = c(-100, 50),
     main = "Distribution of *mean* RT differences\n(across 1000 replications with 10000 participants)")
```

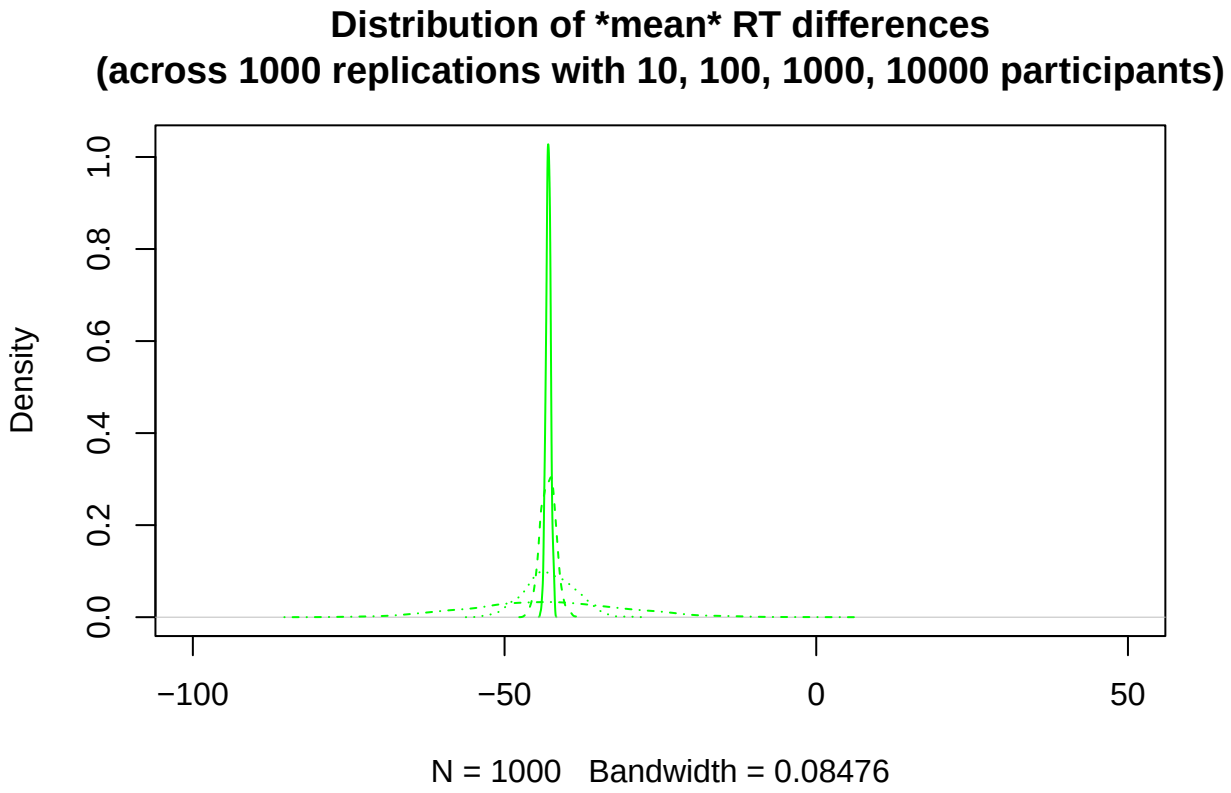


This shows—under the assumptions we've been making about the distribution of RTs for high and low frequency objects—how likely/unlikely the difference in RTs is to be zero in an sample of 10000 observed pairs of RTs (one pair each for each pair of high and low frequency words). Very unlikely!

How do you think this graph would change if you only had 10, 100, or 1000, instead of 10000 pairs of observations? For this, you can draw 1000 samples of size 10, 100, or 1000 data points (instead of 10000). Conveniently, we set up the my_sample() function above so that it allows us to specify the number of observations per sample:

```r
means.RT.difference.1000 = replicate(1000, mean(my_sample(n = 1000)))
means.RT.difference.100 = replicate(1000, mean(my_sample(n = 100)))
means.RT.difference.10 = replicate(1000, mean(my_sample(n = 10)))
```

```r
# let's add that information to the same plot, as a dashed line
plot(density(means.RT.difference), col = "green", xlim = c(-100, 50),
     main = "Distribution of *mean* RT differences\n(across 1000 replications with 10, 100, 1000, 10000
lines(density(means.RT.difference.1000), col = "green", lty = 2)
lines(density(means.RT.difference.100), col = "green", lty = 3)
lines(density(means.RT.difference.10), col = "green", lty = 4)
```



A wee lil' closing exercise

Now try to do the same for the distribution of the *standard deviation* (rather than means). Compare what 1000 repetitions would tell you about the distribution of the standard deviation if you had 10, 100, 1000, vs. 10000 samples in each repetition.