Sampling based approaches—a wee lil' exercise

T. Florian Jaeger

Summer 2023

This document

This document is part of a series of lectures that introduce students to basic data wrangling, plotting, and general approaches to data analysis. Please see the overview lecture. This and all other lectures are created in R markdown. R markdown combines text with R code, allowing us to see the code and its output, embedded within the text describing the code. If you have the original R markdown file (file extension .Rmd), you can 'knit' the document into an HMTL, PDF, or Word file.

Two approaches to statistics

We can distinguish between **two broad approaches** to the statistical inferences scientists try to draw about the world from the samples they obtain in their experiments. The two approaches are mutually compatible and complement each other. To illustrate the two approaches, imagine an experiment in which subjects see a picture of an object and have to decide as quickly and accurately whether the picture depicts something animate (a living thing) or not. The experimenter collects reaction times (RTs) for two conditions: one in which subjects see an object that is common in everyday life ("high frequency") and one in which the object is rarely seen in everyday life ("low frequency"). We hypothesize that responses in the high frequency condition are faster than responses in the low frequency condition. The experimenter collects 10,000 observations in each condition. How do we quantify whether the reaction times in the two conditions differ?

Analytic approach

In the **analytic** approach, we use mathematical knowledge about distributions. 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. The standard error of the mean difference, which we might use to construct a confidence interval around the mean difference to see whether it's different from zero, is the standard deviation of the difference divided by the square root of the number n of observations that went into the mean minus 1.

So, if we are willing to assume (for now) that RTs are normally distributed we can calculate the mean difference, standard deviation of the difference, and standard error of the difference by simply measuring the means and standard deviations of the RTs in each of the two conditions of the experiment. For example, if the experiment found a mean of 250 ms and standard deviation 21 ms for the high frequency objects, and a mean of 293 ms and standard deviation 32 ms for the low frequency condition, then the difference between the two RT distributions would have a mean of 250 - 293 = `-43`, standard deviation of sqrt(21^2 + 32^2) = `38.2753184`, and a standard error of the mean of `38.2753184` / sqrt(10,000 - 1) = `0.3827723`. Why does the standard error of the mean follow this formula? One way to find out is to (re)read a good introduction to statistics. Another way to understand the formula is through the second approach to statistics.

Sampling-based approach

The second approach to statistics is the **sampling-based approach** that this exercise focuses on. It's a powerful approach that let's us understand statistical concepts, models, and our data by simulating outcomes under different assumptions. But before we turn to that, let's go through the above example. To apply the sampling-based approach to the example experiment, we repeatedly draw samples from the distributions of RTs for the high and low frequency objects and then calculate the difference between each pair of samples. "Drawing samples" here simply refers to taking random samples from the distribution. For example, here is how we would draw 20 samples from the RT distribution for infrequent objects with mean=250 and SD=21 in R:

```
# rnorm stands for taking a *r*andom sample from the *norm*al distribution
# (the functions for random sample generators in R always start with "r".
# e.g., rlnorm(), rgamma(), rstudent(), rbinom(), etc.). You can get help on
# any of those functions by entering ?function_name (e.g., ?rnorm) into the R
# console.
rnorm(n = 20, mean = 250, sd = 21)
```

```
##  [1] 283.9787 250.2145 257.5782 253.6386 264.5047 257.3084 239.6194 266.2056
##  [9] 219.7964 227.4464 254.9436 224.1567 260.2172 232.8734 263.7713 216.3638
## [17] 285.9608 278.0553 254.1471 244.1416
```

Returning to the example, the RT differences from our samples form a distribution, and we can get the mean RT difference and standard deviation of that distribution. By repeating this thought experiment, we can even look at the distribution of the mean RT differences across experiments. We'll see that the standard deviation of the distribution of these mean differences across replications of the thought experiment is the same as the standard error of the mean in the parametric example. In other words, the standard error of the mean is simply the expected standard deviation of the distribution of sample means (also called the *sampling distribution of the mean*) that would result from replicating the experiment many times.

More generally, we can obtain any quantity from the distribution—even more complex ones—by repeatedly sampling from it. 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 (which is quite common) or when *we* are not sure about the analytic solution. It can also be a powerful tool to understand complex modela—e.g., by sampling responses from the model and looking at the type of distribution that model predicts (Bayesians call this the "posterior predictive").

But let's take a step back. In this exercise, you'll be going through a simple example of the sampling-based approach.

Building intuitions about the sampling-based approach

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)
```

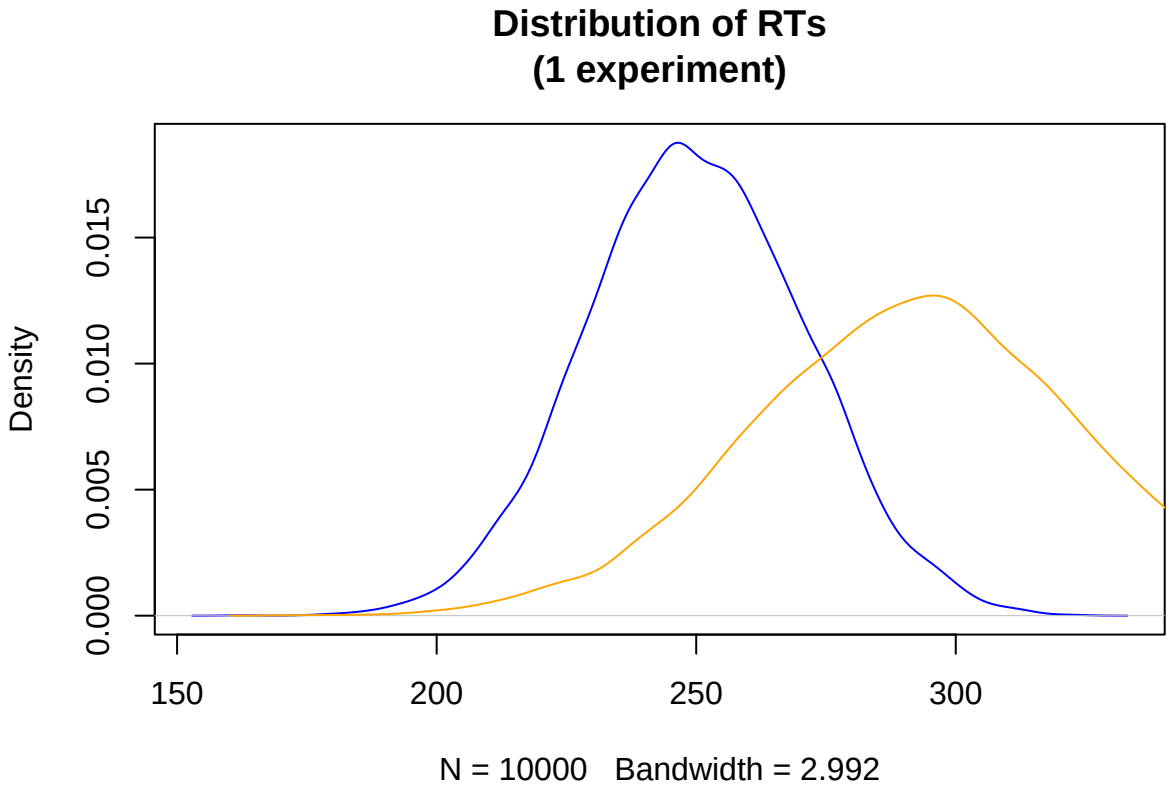
```
RT.low = rnorm(n = 10000, mean = 293, sd = 32)

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

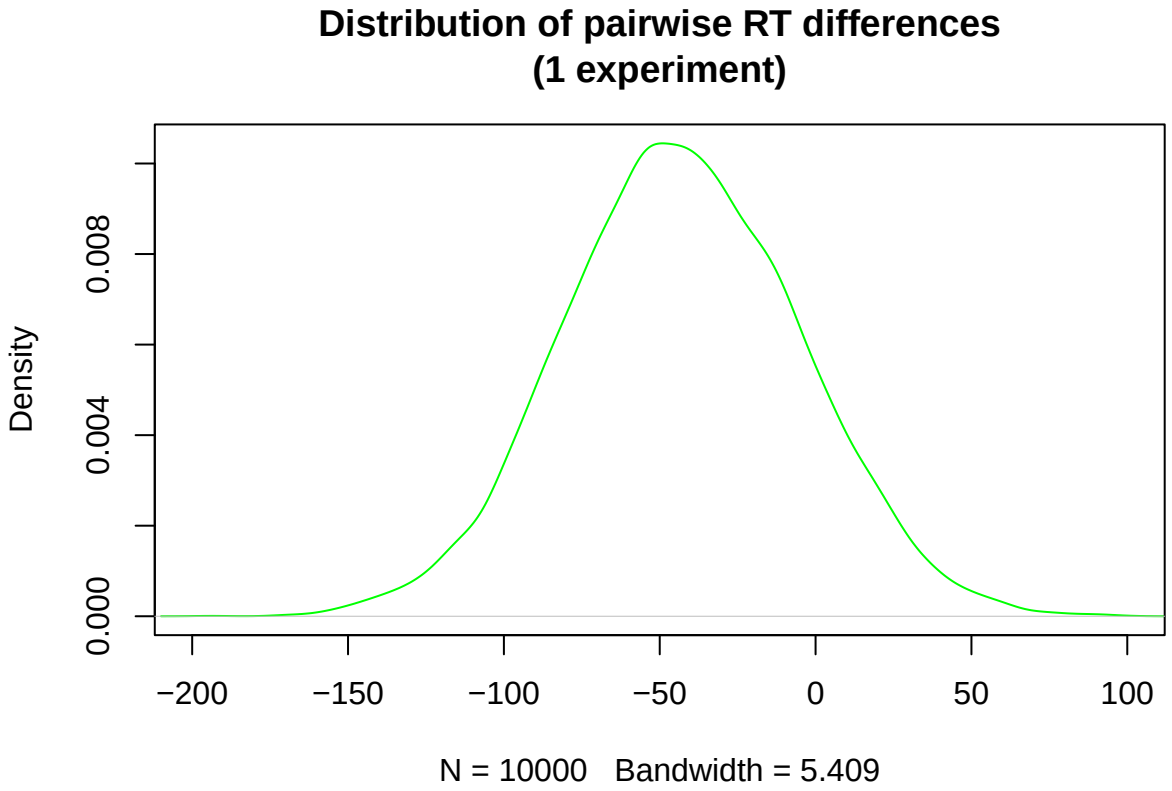
Let's look at the first 10 of these RT differences

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

```
##  [1] "246.990439652028; 280.634201886915; -33.6437622348878"
##  [2] "227.256842872196; 295.524678472252; -68.2678356000557"
##  [3] "255.099065761956; 259.809399133042; -4.7103333710858"
##  [4] "240.030171919694; 311.625417493179; -71.5952455734846"
##  [5] "279.794499819066; 272.109216034071; 7.6852837849948"
##  [6] "240.859873244425; 249.586312854202; -8.72643960977712"
##  [7] "281.091395412063; 300.976336902632; -19.8849414905682"
##  [8] "209.809226780628; 262.457632068644; -52.6484052880168"
##  [9] "227.154884513835; 344.548639247869; -117.393754734034"
## [10] "227.581280722185; 333.436211672531; -105.854930950346"
```

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] -42.83502
```

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

```
## [1] 37.9196
```

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. Let's demonstrate that by trying to understand some of the motivation behind the formula for the standard error of the mean.

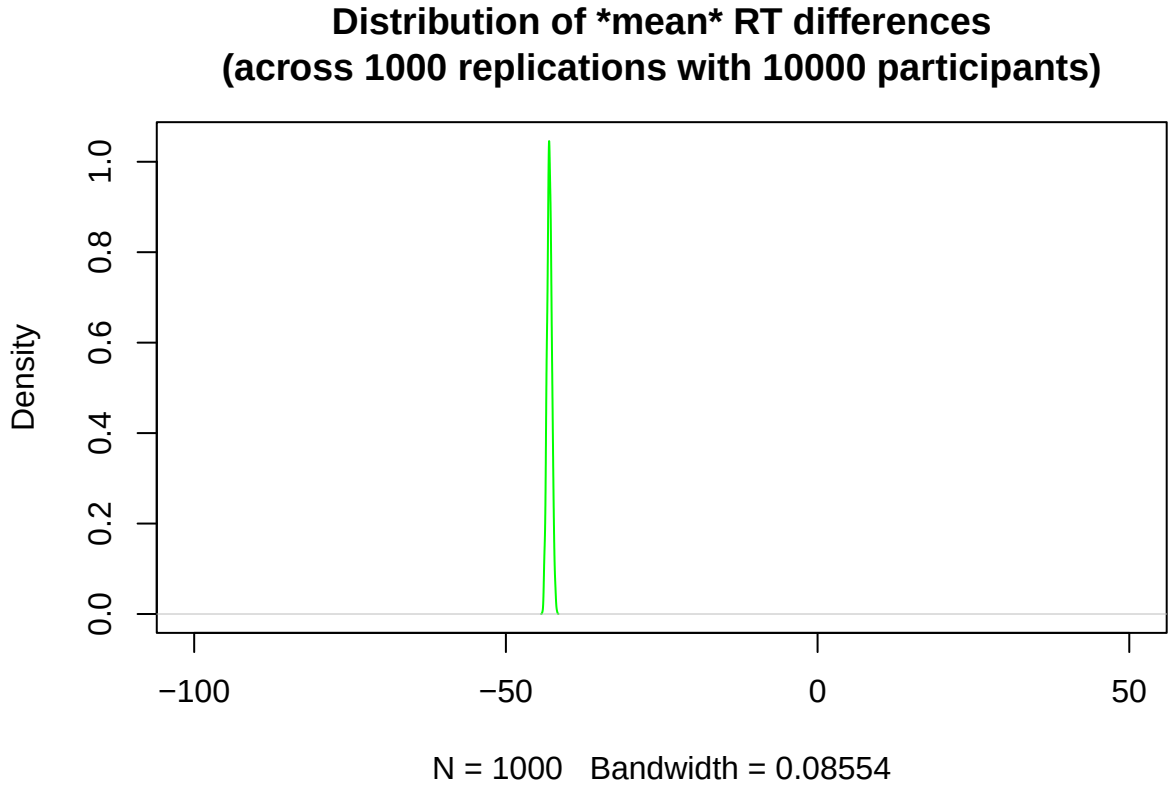
Repeated sampling: continuing our example

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] -42.83249 -42.96775 -43.33372 -43.63785 -43.15576 -43.03019 -42.84457
##  [8] -42.73031 -43.16911 -42.93450 -42.93316 -43.14978 -43.22792 -43.15088
## [15] -43.29324 -43.51943 -42.13923 -43.24573 -43.13773 -43.33393
```



This shows—under the assumptions we've been making about the distribution of RTs for high and low frequency objects—what *distribution* of average RT differences we would expect if we repeated the experiment over an over again. Put differently, the figure above shows us the expected distribution of mean differences based on 1000 simulated instances of the experiment. We can quantify the uncertainty that an experimenter should have about the true mean after those 1000 experiments. Let's calculate the standard deviation of the mean differences in the above figure:[1]

```r
sd(means.RT.difference)
```

```
## [1] 0.3783961
```

This number should look familiar. It's quite similar to the analytically calculated standard error mentioned at the top of this document![2] So, the standard error of a sample mean is simply the standard deviation of

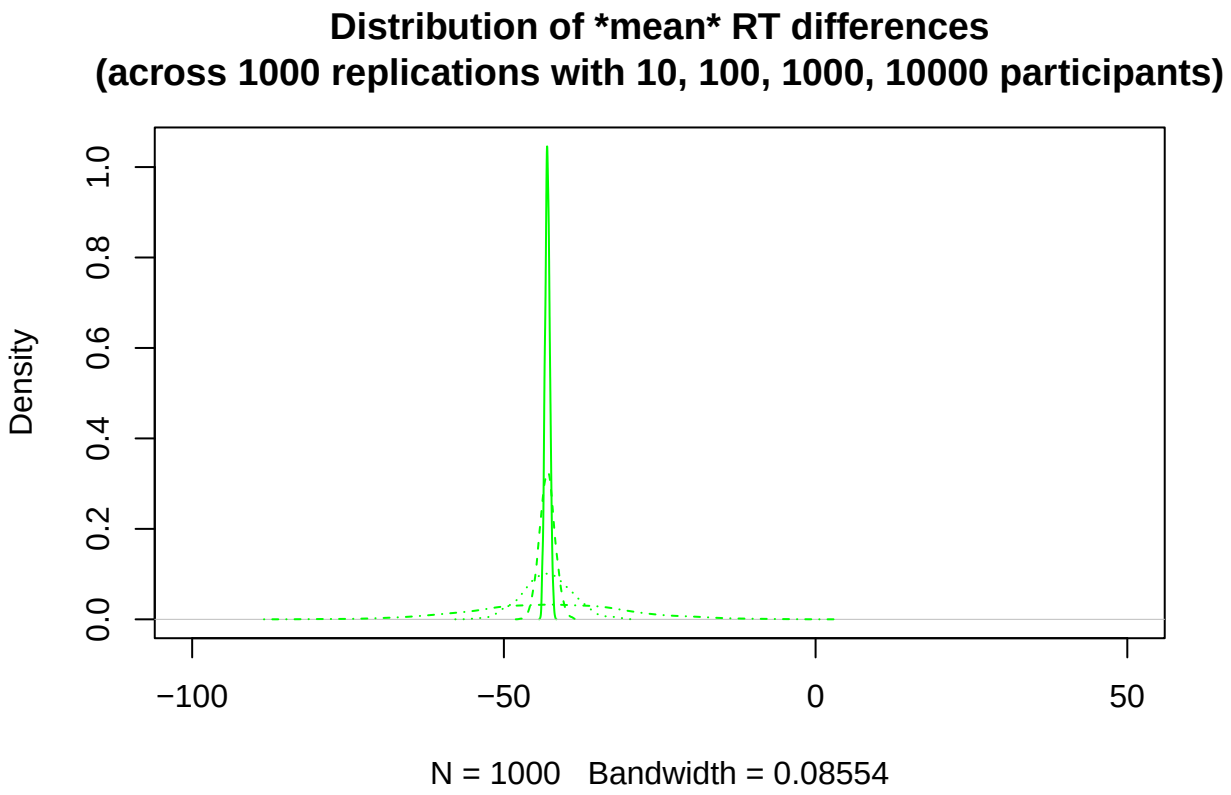
[1] Note that this number will not systematically increase if we run more instances of the experiment (try it out!). Rather, increasing the number of thought experiments from 1,000 to e.g., 100,000 will increase the *accuracy* of this *estimate* of the true standard deviation of the underlying mean difference in RTs.

[2] One reason the number is not entirely identical is that our analytical estimate simply assumed that two means and standard deviations. But each repeated sample of our thought experiment has slightly different means and standard deviations for the two experimental conditions. Another reason is that we only repeated the experiment 1000 times. If we repeat it more often, that will increase the accuracy of the estimate of the standard deviation of the mean differences. Give it a try! Change the code

the means that we would obtain by repeating that experiment many times.

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 n = 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:

```
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)))
```



Notice how the distribution is wider (and the standard deviation of the means larger) when each of the repeated experiments has fewer observations. If an experiment has fewer observations, that increases the experimenter's uncertainty about the true mean (or here the true mean RT difference), as measured by the standard error of the mean. The analytic approach tells us that this standard error is sd(RT difference) / sqrt(n - 1). In the sampling-based approach, we estimate the standard error of the mean by taking the standard deviation of the mean RT differences across repeated instances of the experiment, sd(*mean* RT difference):

```
sd(means.RT.difference.1000)
```

```
## [1] 1.239688
```

```
sd(means.RT.difference.100)
```

```
## [1] 3.958533
```

to run, e.g, 20,000 instances of the experiment.

```r
sd(means.RT.difference.10)
```

```
## [1] 12.0514
```

A wee lil' closing exercise for the curious

If you'd like to try out the simulation-based approach a bit more, try to implement the repeated sampling for the distribution of the *standard deviation* of the RT differences (rather than mean of the RT differences). 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.