1. A group of researchers is running an experiment over the course of 30 months, with a single observation collected at the end of each month. Let $X_1, ..., X_{30}$ denote the observations for each month. From prior studies, the researchers know that

$$X_i \sim f_X(x),$$

but the mean μ_X is unknown, and they wish to conduct the following test

$$H_0 : \mu_X = 0$$
$$H_a : \mu_X > 0.$$

At month k, they have accumulated data $X_1, ..., X_k$ and they have the t-statistic

$$T_k = \frac{\bar{X} - 0}{S_k/\sqrt{n}}.$$

The initial plan was to test the hypotheses after all data was collected (at the end of month 30), at level $\alpha = 0.05$. However, conducting the experiment is expensive, so the researchers want to "peek" at the data at the end of month 20 to see if they can stop it early. That is, the researchers propose to check whether t_{20} provides statistically discernible support for the alternative. If it does, they will stop the experiment early and report support for the researcher's alternative hypothesis. If it does not, they will continue to month 30 and test whether t_{30} provides statistically discernible support for the alternative.

(a) What values of t_{20} provide statistically discernible support for the alternative hypothesis?

```
# (a) t-val for statistically discernible support for t20
# gives t val at 95th percentile
(val_t20 <- qt(0.95, df=19))

## [1] 1.729133
```

To determine whether the researchers should stop the experiment at month 20, we calculated the critical value of the one-sided t-test at the 5% significance level with 19 degrees of freedom (since $n = 20$). This gave a critical value of 1.729, so if $t_{20} > 1.729$, the researchers would reject H_0 and stop early.

(b) What values of t_{30} provide statistically discernible support for the alternative hypothesis?

```
  # (b) t-val for statistically discernible support for t30
  # gives t val at 95th percentile
  (val_t30 <- qt(0.95, 29))

## [1] 1.699127
```

If the experiment continues to month 30, the researchers use the full sample to compute a new test statistic and compare it to the critical value for a t-test with 29 degrees of freedom (since $n = 30$). This gave a critical value of 1.699, so if $t_{30} > 1.699$, they would reject H_0 at that point.

(c) Suppose $f_X(x)$ is a Laplace distribution with $a = 0$ and $b = 4.0$. Conduct a simulation study to assess the Type I error rate of this approach.
Note: You can use the `rlaplace()` function from the `VGAM` package for `R` (Yee, 2010).

```
# (c) simulation to estimate type 1 error

library(VGAM)

## Loading required package:  stats4
## Loading required package:  splines

simulations <- 10000
alpha <- 0.05

type_1_error_count <- 0

for (i in 1:simulations) {
  # generate data from Laplace
```

```r
  data <- rlaplace(30, location=0, scale=4)

  # perform t test
  t20_result <- t.test(data[1:20], mu=0)
  t20 <- t20_result$statistic

  t30_result <- t.test(data, mu=0)
  t30 <- t30_result$statistic

  # check if we would reject the null at month 20
  # do the t-stats exceed the corresponding critical values?
    # if they do, reject the null
  if (t20 > val_t20) {
    type_1_error_count <- type_1_error_count + 1
    # if this doesnt reject, then we check t30
  } else if (t30 > val_t30) {
    type_1_error_count <- type_1_error_count + 1
  }
}

# estimate type 1 error rate
(type_1_error_rate <- type_1_error_count / simulations)

## [1] 0.0756
```

To examine the impact of this strategy, we simulated 10,000 experiments under the null hypothesis, where data were drawn from a Laplace distribution with mean zero and scale 4. In each simulation, we computed t_{20}. If it was significant, we stopped and counted it as rejection. Otherwise, we computed t_{30} and checked again. This procedure gave a type one error rate of approximately 0.073, or 7.3%, which is noticeably higher than the intended 5%. This confirms that checking twice without adjusting for it increases the overall chance of making a Type I error.

(d) **Optional Challenge:** Can you find a value of $\alpha < 0.05$ that yields a Type I error rate of 0.05?

2. Perform a simulation study to assess the robustness of the T test. Specifically, generate samples of size $n = 15$ from the Beta(10,2), Beta(2,10), and Beta(10,10) distributions and conduct the following hypothesis tests against the actual mean for each case (e.g., $\frac{10}{10+2}$, $\frac{2}{10+2}$, and $\frac{10}{10+10}$).

   (a) What proportion of the time do we make an error of Type I for a left-tailed test?

   (b) What proportion of the time do we make an error of Type I for a right-tailed test?

   (c) What proportion of the time do we make an error of Type I for a two-tailed test?

   (d) How does skewness of the underlying population distribution effect Type I error across the test types?

References

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32(10):1–34.