1. **Cancer and Group Therapy**. Researchers randomly assigned metastatic breast cancer patients to either a control group or a group that received weekly 90-minute sessions of group therapy and self-hypnosis. The group therapy involved discussion and support for coping with the disease. The goal of the experiment was to see whether the latter treatment improved the patients' quality of life at the time, but a followup study on these patients collected data on the number of months of survival after the beginning of the study[1].

   You can access the data from this study using the following code.

```
library(tidyverse)
cancer <- read.csv("https://stat158.berkeley.edu/spring-2026/data/breast-cancer/breast-cancer.csv")
```

   `GROUP` is the original group assigned to the subject. `SURVIVAL` is the survival time in months from the beginning of the study. At the time the survival data was collected (10 years later), some subjects were still alive. They are flagged in the `CENSOR` column. Data (in this case survival time) that can only be known up to some bound is called *censored*.

   a. Choose an effective way to visualize the data and present the plot. Comment on what you see.
   b. Is there evidence of an effect of the group therapy on survival time?
      i. Answer this using randomization inference and two different test statistics: the difference in means and the difference in medians[2].
      ii. Answer this using model-based inference using a two-sample t-test using the pooled estimate of the standard deviation[3].
      iii. Compare the methods by first describing the different ways they conceive of randomness and then comparing the results and what they say about the research question.
   c. When conducting Randomization Based Inference, as you did above, what group of units does the inference directly apply to? Said another way, the parameter under study is a function of which group of units?
   d. What problems do you see with having censored data and what problem could it *potentially* create for your analysis? Do you think it has affected your conclusions for this particular dataset?

[1] Data from Q 4.31 in *Statistical Sleuth*.

[2] You can augment the `rand_stats()` function to take `stat = c("diff in mean", "diff in medians")` as an argumnent and then copy and modify the `if` block to calculate the appropriate statistic when `stat == "diff in medians"`.

[3] This should be a review from your course in Statistical Inference. If you need a brush up, see https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm and `?t.test` in R.

2. **Statistical Errors in Testing**. Consider the study of whether group therapy can have an effect on breast cancer survival.
   a. What would constitute a type I error?
   b. What would constitute a type II error?
   c. Why do statistical errors in testing take place? Is there any way to completely eliminate them?
3. **The Levers of Power**. Consider the study of whether group therapy can have an effect on breast cancer survival.
   a. What is statistical power in this context?
   b. Imagine you were advising the researchers as they were planning their study. What are four different parameters that effect the power that they could consider optimizing? Realizing that they have practical constraints on what they can do, what guidance can you give for selecting values of those parameters?

4. **Anchoring Power Curve (function of effect size)**. A power curve shows, for a particular experiment, the relationship between some parameter in the experiment and the statistical power. In Randomization-Based Inference, they are approximated point-by-point, calculating the power for different values of one parameter while keeping the others fixed.

   Create a power curve that shows the relationship between the effect size (as measured by τ, a "constant shift" ITE) and the power for the Anchoring Experiment. Recall this is a two-tailed test conducted at $\alpha = .05$. Fix the sample size at the values observed in the data. Consider at least five values for τ between 0 and 20 (you're encouraged to consider more values and consider negative values). Put those values of τ along the x-axis and the corresponding values of power on the y-axis.

   Tip: this exercise will require a lot of code unless you wrap stretches of it in functions. See the last slides from class for functions that you're welcome to use (if you'd like more practice, try writing your own first).

5. **Practice with the t.**[4] Create a visualization of each of the following functions. Ensure the axes are labelled in an informative manner and add appropriate titles. Choose x-axis limits that focus on the most interesting part of the functions.[5]
   a. The probability density function (PDF) of the t distribution with 4 degrees of freedom with vertical lines marking the .025 and .975 quantiles of the distribution.
   b. The cumulative distribution function (CDF) of the t distribution with 4 dsegrees of freedom.
   c. The PDF of the non-central t distribution with 4 degrees of freedom and a non-centrality parameter of .3.
6. **Anchoring Model-Based Power Curve (function of effect size)**.
   a. Using the t distribution, create a power curve that shows, for the Anchoring Experiment, the relationship between the effect size (as measured by $\mu_1 - \mu_0$) and the power for a two-tailed test conducted at $\alpha = .05$. Fix the sample size at the values observed in the data and use 400 as your value of σ^2.
   b. At what effect size does the power reach .8?
   c. *Optional Challenge:* Overlay the power curve you created in part a with the power curve you via randomization-based inference in the previous exercise. Do they look similar? If not, can you explain why not?

[4] In R, for the t distribution, the PDF is `dt()`, the quantile function is `qt()`, and the CDF is `pt()`. There are analogously named functions for all common named probability distributions, e.g. `dbinom()` and `pnorm()`.

[5] An easy way to plot functions in R is to create a data frame with a column `x` that has a sequence of many evenly-spaced values along the x-axis (try 100 values). You can then add additional columns for each function of `x` that you wish to visualize. You can plot a pair of those columns against one another and connect the points with a line.