1. **The Secret to Longevity**. Scientists are interested in whether the energy costs involved in reproduction affect longevity[1]. In this experiment, 125 male fruit flies were divided at random into five sets of 25. In group 1, the males were kept by themselves. In groups 3 and 5, the males were supplied with one or eight receptive virgin female fruit flies per day, respectively. In groups 2 and 4, the males were supplied with one or eight unreceptive (pregnant) female fruit flies per day, respectively. Other than the number and type of companions, the males were treated identically. The longevity of the flies was observed.

   You can access the data as a csv file at: https://stat158.berkeley.edu/spring-2026/data/fruit-flies/fruit-flies.csv

   a. Visualize the data and interpret what you see. Be sure to comment both on the center of the distributions and their spread.

   b. *ANOVA "by hand"*. Without using any of the built-in commands (`aov()`, `anova()`) create a one-way ANOVA table minus the p-values. You're welcome to use either a base R or `dplyr` approach. It is a bit tedious, however it's useful to cobble together each of the statistics in an ANOVA table just once in your lives. You can check your answers with the output of `aov()` shown below.

```
             Df Sum Sq Mean Sq F value Pr(>F)
factor(trt)   4  11939  2984.8   13.61
Residuals   120  26314   219.3
```

   c. State your conclusion about the scientific question using Model-based inference. Be clear about the data generation model your approach relies upon and the null hypothesis that you're evaluating. To calculate the p-value, use the `pf()` function and then verify it using the `aov()` function[2].

   d. State your conclusion about the scientific question using Randomization-based Inference. Be clear about the data generation model your approach relies upon and the null hypothesis that you're evaluating. For your test statistic, use the F-statistic that you built the code for from part a above.

   e. How do the results of the model-based and randomization-based inference compare? Do they lead to the same conclusion? Do they rely on the same assumptions? Which do you think is more appropriate for this problem?

2. **More Comparisons for Longevity**. In the experiment regarding the fruit flies, the ANOVA analysis only tests whether there were any differences between the group means. Suppose there was further interest in whether the 1 pregnant group was different from the 1 virgin group (for questions a and b, restrict the data set to these two groups).

   a. Perform a randomization test to compare these two groups using the t-statistic. Assume the two groups have the same variance (revisit your notes on the t-test and the pooled estimate of the SE).

[1]Question from Problem 3.2 in Oehlert textbook, original data from Hanley and Shapiro (1994), ``Sexual activity and the lifespan of male fruitflies: a dataset that gets attention," *Journal of Statistics Education 2*.

[2]Recall all named probability distributions have a CDF in R with the prefix `p`. The F-distribution is no exception, so you can use `pf()` to calculate the p-value for your ANOVA table. The arguments to `pf()` are the F-statistic, the degrees of freedom for the numerator, and the degrees of freedom for the denominator. Recall a test using the F statistic is a right-tailed test. The standard anova output is available with `summary(aov(y ~ x, data))` (be sure your `x` is a factor vector).

b. Repeat the randomization test but this time use an F-statistic. How do the results of this analysis this compare to the results in part (a)?

c. Calculate an ANOVA table for this comparison and report the value of $\sqrt{MS_{residuals}}$, which is an estimate of σ. How does that compare to the estimate of the standard deviation that you used for the t-statistic (s_p)?

You are welcome to use built-in R commands for this question, if appropriate.

3. **Poison**. Medical researchers investigated how to combat the effects of certain toxic agents (indeed poison, not Poisson!). For this, they randomly assigned 48 to a group where they are treated with a poison and a drug to counteract the poison. The survival time of the animal was recorded as the response of interest[^death]. Three poisons and four drugs were considered, and four mice were assigned to each poison/drug combination. The responses are recorded in hours.

   You can access the data as a csv file at: https://stat158.berkeley.edu/spring-2026/data/poison/poison.csv

   a. Create an interaction plot of the data and interpret it.
   b. Fit a two factor model with interactions and
      - i. for each component of the model (main effect of poisons, main effect of drugs, interaction effects), calculate the value of observed F statistics and the corresponding p-values. You may calculate the p-values using either randomization-based inference or model-based inference (or both)[3].
      - ii. draw conclusions about differences among poisons, differences among treatments, and interactions between them. Is this a setting where it is important to estimate the interaction effects?

[3] Note the interpretation of p-values associated with a test on interaction terms is different for RBI than MBI due to different null hypotheses. To test a null hypothesis analogous to the MBI that $\gamma_{j,k} = 0 \forall j, k$ using RBI requires a more sophisticated shuffling method such as the Freedman-Lane Residual Permutation Test.

4. **Growing Herbs**. Botanists studying optimal greenhouse conditions for a fast-growing herb ran a fully-crossed three-way factorial experiment. They manipulated light intensity (low, medium, or high), watering frequency (low, medium, or high), and soil type (sandy or loam). Each of the $3 \times 3 \times 2 = 18$ treatment combinations was replicated twice, yielding $n = 36$ plants. The response is shoot dry weight (grams) measured after four weeks.

   You can load the data directly in R with the following code chunk:

```
plants <- data.frame(
  weight = c(12.7,14.2,15.7,12.6,16.4,17.2,15.0,17.4,20.5,
             11.9,15.2,16.4,12.9,17.9,18.7,15.9,18.1,25.7,
             11.3,13.4,15.5,12.0,16.6,18.0,15.6,17.6,22.1,
             14.8,13.9,18.3,14.8,17.8,19.2,16.7,17.9,26.4),
  light  = gl(3, 1, 36, labels = c("low", "medium", "high")),
  water  = gl(3, 3, 36, labels = c("low", "medium", "high")),
  soil   = gl(2, 9, 36, labels = c("sandy", "loam"))
)
```

   a. Create a *main effects plot* for each of the three factors — light intensity, watering frequency, and soil type — and interpret each one. Which factor appears to have the largest main effect on shoot dry weight?

   b. Construct *three two-way interaction plots* (one for each pair of factors: light × water, light × soil, water × soil) and interpret them. Do any of the pairs show evidence of an interaction?

   c. Construct a *three-way interaction plot* by plotting mean shoot dry weight against one factor on the x-axis, a second factor mapped to color and line type, and using `facet_wrap()` on the third. Interpret what you see: is there evidence that a two-way interaction changes in character across levels of the third factor?

   d. Fit a full three-way ANOVA model using `aov(weight ~ light * water * soil, data = plants)`. Report the ANOVA table and, for each term in the model, state whether you would conclude the effect is present.

5. **Estimation Three Ways**. We laid out the additive model for a two-way factorial with no interactions as

$$Y_i = \mu + \alpha_{j(i)} + \beta_{k(i)} + \epsilon_i.$$

   The most commonly used estimators for these parameters are as follows.

$$\hat{\mu} = \hat{\bar{Y}}$$
$$\hat{\alpha}_j = \hat{\bar{Y}}_j - \hat{\bar{Y}}$$
$$\hat{\beta}_k = \hat{\bar{Y}}_k - \hat{\bar{Y}}$$

   There are three different reasons that these estimates are sensible:

   a. They provide unbiased estimates of the corresponding causal parameters (shown in class).

b. These estimates are the least squares estimates. That is, they are the values that minimize

$$\sum_{i=1}^{n} \left(Y_i - \left(\mu + \alpha_{j(i)} + \beta_{k(i)}\right)\right)^2$$

c. Under a generative model assumption that $\epsilon_i \sim N(0, \sigma^2)$, these are the maximum likelihood estimates.

Pick either b. or c. and brush off your calculus skills to demonstrate that it is true. If you haven't seen maximum likelihood estimation before, stick to b.