1. **Two-sample test:** One of the ways that small businesses advertise is to put a flyer (a paper advertisement) on the windshield of every car in a large shopping center parking lot. A study was carried out to compare the response rate to advertisements printed on different colored paper. Twelve shopping center parking lots were randomly assigned to receive orange or red ads. For each parking lot the response variable is the number of phone calls received in the next 24 hours after the flyers were distributed multiplied by 100 and divided by the number of flyers placed. We’ll call this the response rate. The response rates are summarized below:

<table>
<thead>
<tr>
<th></th>
<th>Orange</th>
<th>Red</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>mean</td>
<td>3.1</td>
<td>2.7</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.85</td>
<td>0.65</td>
</tr>
</tbody>
</table>

(a) Is this study an experiment? Explain.

(b) Define the population of interest, the experimental units, the treatment and the response.

(c) An alternative (easier) design might use the individual car as a unit. Then you could just use each color in one parking lot and take the response from each car as 0 (if nobody called) or 1 (if somebody called) – you’d have to identify each flyer with a code or something. Answer the question from part (b) for this study design. Explain any concerns you might have about this design.

(d) The effect size is a measure of the difference in two population means (i.e., it measures the effect of the treatment). There are various definitions for effect size depending on the study. For a two sample study the effect size is usually defined as \( \mu_2 - \mu_1 \) (if the units of measurement are meaningful) or \( (\mu_2 - \mu_1) / \sigma \). These can be estimated by plugging in sample estimates of the relevant quantities. Using the second definition what is the effect size for this study? (Note: Interpretation of effect sizes depends on the field. An effect size of 1.0 would be a very large effect, shifting the mean of the population by 1.0 standard deviation.)

(e) Show that the usual pooled two-sample t-test statistic can be written as the effect size times some function of sample size.

(f) Carry out a test of the null hypothesis that the two colors are equally effective. Calculate the p-value and summarize your conclusion in one or two sentences.

(g) Suppose the study is repeated with four times as many parking lots (n=28 in each group). You get EXACTLY the same mean and s.d. Would the effect size estimate change? Would the p-value change? Interpret the answers to these two questions for the investigator (e.g., did the second study work better?)

2. **Two-sample confidence interval:** In one weight-loss study 89 sedentary men were randomly assigned to either a special diet or exercise for a year. Forty-two men were placed on a diet and they lost an average of 8.2 kg with a standard deviation of 3.8 kg. The other 47 men were put on an exercise program and they lost an average of 6.3 kg with a standard deviation of 4.1 kg.

(a) Use the two sample t procedure to find a 95% confidence interval for the difference between the mean weight loss via diet and the mean weight loss via exercise.

(b) Give a 90% CI for the difference in means? Give a 99% CI for the difference in means?

(c) Use the CI results to tell what p-value would be obtained in a two-sample t-test. (Hint: You can’t give a precise p-value based only on the CIs, only an interval. You can do the test to check but should report the answer to the question that is asked!)

(d) It is pointed out that the distribution of weight loss values for the men don’t appear to follow a normal distribution. Is this a concern? Explain why or why not.

3. **Theory – Power calculation** A key advantage of the model-based approach to inference is that the model can be used to plan studies. This problem revisits the two-sample power calculation discussed briefly in class. Suppose that we will be collecting data from two populations and are willing to assume that the samples \( Y_{11}, \ldots, Y_{1n} \) are iid \( N(\mu_1, \sigma^2) \) and the samples \( Y_{21}, \ldots, Y_{2n} \) are iid \( N(\mu_2, \sigma^2) \) as required by the model. To keep things simple here let’s assume \( \sigma \) is known and that therefore we can use normal (rather than t) tables in carrying out our tests.

(a) We wish to test the null hypothesis that \( \mu_2 = \mu_1 \) versus the alternative hypothesis that \( \mu_2 > \mu_1 \). We are going to use a size \( \alpha \) test (i.e., we will reject \( H_0 \) if the p-value is less than \( \alpha \)). Show that the usual z-test rejects \( H_0 \) if \( \bar{Y}_2 - \bar{Y}_1 > c \) and identify \( c \) (it should involve \( \sigma, n \) and \( z_{1-\alpha} \)).
Hospital stays

Theory: Effect of non-independence

4. **Hospital stays** – Please perform this analysis using R (or your other preferred software). The Study on the Efficacy of Nosocomial Infection Control (SENIC) was a study of U.S. hospitals focused on studying hospital-acquired (nosocomial) infection rates and the factors associated with them. (The data set is further described in Appendix C in the text.) For this homework we focus on whether a hospital being associated with a medical school leads to higher infection rates. The data file senic.csv provides information about 113 hospitals. The data file contains 12 variables:

- id = hospital identification number
- stay = average length of stay per patient (in days)
- age = average patient age (in years)
- inf = infection risk (infections per 100 patients)
- cult = routine culturing ratio
- xray = routine x-ray ratio
- beds = number of beds
- m = medical school affiliation (1=yes, 2=no)
- r = region (1=NE, 2=NC, 3=S, 4=W)
- pat = average number of patients
- nur = average number of full-time equivalent nurses
- facil = percent of facilities/services offered

(a) We focus on comparing inf across the two groups having m=1 and m=2. Do the assumptions required for an analysis based on the pooled t-test appear to be satisfied? Explain and provide supporting evidence. (R Hints: You can operate on a selected subset of a vector by incorporating it as a reference. Thus hist(inf|m==1)) will create a histogram of the infection rates using only those hospitals for which m is equal to one. Another useful function for checking the assumptions is the qqnorm function.)

(b) Regardless of your answer in (a), carry out a pooled two-sample t-test of the hypothesis that the mean infection rate for hospitals with a medical school is the same as the mean infection rate for hospitals without one. (R Hints: In class we used the t.test function in one of its formats, t.test(y1,y2) are the data from group one and y2 are the data from group two. You can also use the alternate format t.test(y ~ x) where y is all of the data and x defines the groups.)

(c) Report a 95% confidence interval for the difference in mean infection rates.

(d) Would this comparison of the two types of hospitals be considered an experiment or an observational study? Explain. How does this impact the interpretation of the results you found?

(e) Write a 2-4 sentence summary of the results of your investigation. (Address what you were studying, what you found, and any concerns)

5. **Theory: Effect of non-independence**

(a) Assume that $Y_1, \ldots, Y_n$ are independent identically distributed (iid) random variables with mean $\mu$ and $\sigma^2$. Show that $\bar{Y}_n = \frac{1}{n} \sum_i Y_i$ has mean $\mu$ and variance $\sigma^2/n$.

(b) Suppose that $Y_1, \ldots, Y_n$ are identically distributed random variables with mean $\mu$ and variance $\sigma^2$ but they are no longer independent. Instead assume that $Y_i$ and $Y_j$ have correlation $\rho$ (equivalent to covariance $\sigma^2 \rho$) for all pairs $i$ and $j$; all observations are correlated with each other. Find $E(\bar{Y}_n)$ and $\text{Var}(\bar{Y}_n)$ in this case.

(c) Let’s study the impact of this correlation on the variance of $\bar{Y}_n$:

i. Compute the variance of $\bar{Y}_n$ if $n = 100$ and $\rho = 0.1$. How many independent observations provide the same information as these 100 correlated observations? This is sometimes known as the effective sample size.

ii. Compute the variance of $\bar{Y}_n$ if $n = 100$ and $\rho = 0.01$. How many independent observations provide the same information as these 100 correlated observations?

iii. Compute the variance of $\bar{Y}_n$ if $n = 100$ and $\rho = -0.05$. How many independent observations provide the same information as these 100 correlated observations? (Warning: Be careful. There is something not quite right here.)

NOTE: You will need to use formulas from probability/statistics theory ($E(X+Y) = E(X) + E(Y)$ and $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X,Y)$).