1. We are planning a study in which our goal will be to accurately estimate a treatment effect, expressed as the difference in mean responses between treated and untreated subjects. Our goal is to have the standard error for our estimate be around 20% of the true value. Based on previous studies, we anticipate the true value to be around 2. We also expect, based on previous research, that the treated subjects’ standard deviation is 1, and the untreated subjects’ standard deviation is 2. Our study will enroll twice as many treated subjects as untreated subjects.

(a) What total sample size should we obtain?

**Solution:** The standard error of the estimated treatment effect should be $0.2 \cdot 2 = 0.4$, thus the variance of the estimated treatment effect should be around 0.16. The estimated treatment effect is

$$\bar{Y}_T - \bar{Y}_U,$$

and the variance of the estimated treatment effect is

$$\sigma_T^2/n_T + \sigma_U^2/n_U = 1/(2n) + 4/n = 9/(2n),$$

where $n$ is the number of untreated subjects, and thus $n + 2n = 3n$ is the total sample size. Thus we need

$$9/(2n) = 0.16,$$

so $n = 9/(2 \cdot 0.16) \approx 28$, so the total sample size should be around 84.

(b) Consider the power for testing the null hypothesis of zero treatment effect, using the sample size you found in part (a), and assuming that the effect size is 2, as supposed. Will this power be greater than 80%? Explain your reasoning, but you do not need to obtain a numerical value for the power.

**Solution:** When $n = 28$, the variance of the estimated treatment effect is $9/56$. If the actual effect size is 1.5, then

$$ET = \frac{2}{\sqrt{9/56}} \approx 4.99$$

Thus the power is


We know from the 68-95-99 rule that 99% of the probability of a standard normal distribution is between -3 and 3. Thus $P(Z > -3)$ is greater than 0.99. So the power is far greater than 80%.
2. Consider the following descriptions of graphs of $y$ (vertical axis) against $x$ (horizontal axis). Match the descriptions of the graphs to the actual graphs below. Each description matches exactly one graph. Two of the graphs match none of the descriptions. Write the letters a-d in the graphs as appropriate.

(a) We observe independent and identically distributed data $X_1, \ldots, X_n$, with mean $\mu$ and variance $\sigma^2$, and use it to form a 95% confidence interval for $E X$. $\sigma$ is the standard deviation that is used when forming the confidence interval. $y$ is the coverage probability of the confidence interval.

(b) We collect 100 data points, $X_1, \ldots, X_{100}$, where $\text{cor}(X_1, X_2) = x$, $\text{cor}(X_3, X_4) = x$, etc. (and all other pairs of $X$ values are independent). $y$ is the variance of the sample mean $\bar{X}$.

(c) We collect two unpaired, independent samples of data $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. The $X_i$ and the $Y_i$ all have the same variance, which is $x$. $y$ is the expected value of the test statistic for the null hypothesis $E Y = E X$.

(d) We construct a 95% confidence interval for $E X$ based on an independent sample of size $x$. $y$ is the coverage probability of this interval.

3. A public health researcher is interested in estimating the level of alcohol consumption among residents in a certain community. Since the community is large, the researcher is planning to divide the community into 200 geographic regions consisting of approximately 100 residents each, then select 10 of the 200 regions at random. The researcher will then obtain data on 10 subjects from each of the 10 selected regions, giving a total sample size of 100 subjects.
(a) What type of sampling is being employed in this study?

**Solution:** This is cluster sampling.

(b) The variance of \( \bar{Y} \), the sample mean of the 100 data values, is

\[
\frac{\sigma^2}{n} + \frac{\tau^2}{q}
\]

where \( \sigma^2 \) is the variance among people in the same region, \( \tau^2 \) is the variance among the mean values in the different regions, \( n \) is the total sample size, and \( q \) is the number of selected regions. It is known from previous work that \( \sigma \) is around 0.8. How big may \( \tau \) be so that the width of a 95% confidence interval for the population level of alcohol consumption is at most 0.5?

**Solution:** The width of the CI is 4 times the standard error, so we need 4SE = 0.5, or for the variance of the estimate to equal 1/64. The variance of the estimate is

\[
0.64/100 + \tau^2/10 = 0.0064 + \tau^2/10.
\]

Thus we need

\[
0.0064 + \tau^2/10 = 1/64,
\]

or \( \tau^2 \approx 0.092 \), or \( \tau \approx 0.3 \).

(c) Place the following three possible study designs in order from least variance to greatest variance.

i. select 10 regions, then select 10 subjects within each selected region (as above)
ii. select 20 regions, then select 5 subjects within each selected region
iii. select 10 regions, then select 8 subjects within each selected region

**Solution:**

\[
\text{var}(\text{iii}) > \text{var}(i) > \text{var}(ii).
\]

4. Suppose we are planning a study in which a log odds ratio will be used to assess the association between having high body mass index (BMI) and high blood pressure (BP). We would like to collect enough data so that the standard error of our estimated log odds ratio is 0.3. For planning purposes, we will use the following hypothetical joint distribution for \( X \) and \( Y \):

<table>
<thead>
<tr>
<th></th>
<th>High BP</th>
<th>Normal BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>High BMI</td>
<td>1/7</td>
<td>1/7</td>
</tr>
<tr>
<td>Normal BMI</td>
<td>2/7</td>
<td>3/7</td>
</tr>
</tbody>
</table>

(a) What sample size is needed?

**Solution:** The variance of the log odds ratio for a sample of size \( n \) is

\[
\frac{7}{n} + \frac{7}{n} + \frac{7}{(2n)} + \frac{7}{(3n)} = 119/(6n)
\]
A standard error of $3/10$ is equivalent to a variance of $9/100$. Thus we have

$$\frac{119}{6n} = \frac{9}{100},$$

so $n = 119 \cdot 100/54 \approx 220$.

(b) Construct a joint distribution that has the same odds ratio as the table given above, but that has marginal probability equal to 0.5 for the high BMI group.

Solution: We know that the odds ratio is not changed if we introduce constants $f$ and $g$, as follows:

<table>
<thead>
<tr>
<th></th>
<th>High BP</th>
<th>Normal BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>High BMI</td>
<td>$f/7$</td>
<td>$f/7$</td>
</tr>
<tr>
<td>Normal BMI</td>
<td>$2g/7$</td>
<td>$3g/7$</td>
</tr>
</tbody>
</table>

Thus we get two equations involving $f$ and $g$:

$2f + 5g = 7$ (since the total probability must equal 1)

$2f = 5g$ (so that the marginal probabilities for BMI are equal)

Solving these, we get

$f = 7/4$, $g = 7/10$.

5. An educational researcher is considering whether reducing class size improves teacher performance. A two-year study is proposed in which 100 teachers will be selected, and after one year of teaching, the mean student test score $X_i$ for each teacher will be obtained. Then in the second year, each teacher’s class size will be reduced by five students, and the student test scores $Y_i$ will be obtained at the end of the second year. For simplicity, suppose that the test scores have been scaled so that $\text{var}(X) = \text{var}(Y) = 1$.

(a) Suppose it is believed that the $X_i$ and $Y_i$ values are correlated at level $r = 0.5$. Thus a paired $t$-test is used for the analysis. How big must the treatment effect $ED = EY - EX$ be so that we have 80% power in our study?

Here are some quantiles of the standard normal distribution:

<table>
<thead>
<tr>
<th></th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.28</td>
<td>-0.84</td>
<td>-0.52</td>
<td>-0.25</td>
<td>0.00</td>
<td>0.25</td>
<td>0.52</td>
<td>0.84</td>
<td>1.28</td>
</tr>
</tbody>
</table>

Solution: Let $\bar{D} = \bar{Y} - \bar{X}$. The variance of $D$ is

$$\text{var}(D) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y) = 2(1 - r)$$

Thus the paired test statistic is

$$10\bar{D}/\sigma_D = 10\bar{D}/\sqrt{2(1 - 0.5)} \approx 10\bar{D}.$$
The power is

\[ P(10\bar{D} > 2) = P(10(\bar{D} - ED) > 2 - 10 \cdot ED) = P(Z > 2 - 10 \cdot ED). \]

If we set \( 2 - 10 \cdot ED = -0.84 \), we get

\[ ED = \frac{2.84}{10} \approx 0.284. \]

(b) Subsequently, new information is obtained, suggesting that the correlation between \( X \) and \( Y \) is around zero. What sample size would be needed so that the power for detecting the effect size you found in part (a) is still 80%?

**Solution:** The test statistic is

\[ \sqrt{n}\bar{D}/\sqrt{2}, \]

so the power for detecting an effect of size of 0.284 is:

\[ P(\sqrt{n}\bar{D}/\sqrt{2} > 2) = P \left( \sqrt{n}(\bar{D} - 0.284)/\sqrt{2} > 2 - \sqrt{n}0.284/\sqrt{2} \right) = P(Z > 2 - 0.2\sqrt{n}). \]

We need

\[ 2 - 0.2\sqrt{n} = -0.84, \]

so

\[ n = (2.84/0.2)^2 \approx 202. \]

6. Suppose we are collecting information on personal debts from married couples in the age range 30-35. Our study aims to compare indebtedness for couples with children to indebtedness for couples without children. The population mean indebtedness of a couple with children is \( \alpha_1 \), and the population mean indebtedness of a couple without children is \( \alpha_0 \). However, the indebtedness for couples without children is measured with bias, so that the expected value of a data point from this group is \( \alpha_0 - \delta \). There is no bias when measuring indebtedness for couples with children. Our interest is the groupwise difference \( \alpha_1 - \alpha_0 \).

We sample 50 couples with children and 50 couples without children, and then form the mean difference \( \bar{Y}_1 - \bar{Y}_0 \) (where \( \bar{Y}_1 \) is the sample mean response for the couples with children, and \( \bar{Y}_0 \) is the sample mean response for the couples without children). For simplicity, assume that the variance for all couples is \( \sigma^2 = 1 \).

(a) Suppose we observe a Z-score of 2.5 for the unpaired t-test that compares the means of the two samples. Using sensitivity analysis, determine the largest value of \( \delta \) for which we can conclude that \( \alpha_1 - \alpha_0 \) is significantly different from zero.
**Solution:** The expected value of the test statistic is
\[
\frac{\alpha_1 - (\alpha_0 - \delta)}{\sqrt{1/50 + 1/50}} = 5(\alpha_1 - \alpha_0 + \delta).
\]
Thus we have
\[
2.5 \approx 5(\alpha_1 - \alpha_0) + 5\delta.
\]
We need
\[
5\delta < 0.5
\]
In order for the treatment effect to have a Z-score greater than 2. So
\[
\delta < 0.5 / 5 = 0.1.
\]

(b) Now suppose we form a 95% confidence interval around \(\bar{Y}_1 - \bar{Y}_0\), with the goal of covering \(\alpha_1 - \alpha_0\). What is the coverage probability?

**Solution:** Let \(D = \bar{Y}_1 - \bar{Y}_0\), and let \(s = 0.2\) denote the standard deviation of \(D\). Then the coverage probability is
\[
P(D - 2s \leq \alpha_1 - \alpha_0 \leq D + 2s) = P(-2s \leq D - (\alpha_1 - \alpha_0) \leq 2s)
\]
\[
= P(-2s - \delta \leq D - (\alpha_1 - \alpha_0 + \delta) \leq 2s - \delta)
\]
\[
= P(-2 - \delta / s \leq Z \leq 2 - \delta / s).
\]