Due in lab on Friday, October 1st

1. Suppose we are interested in estimating a quantity $\mu$, and we can obtain independent measurements $X_i$ with $EX_i = \mu$. We intend to use $\bar{X}$ to estimate $\mu$, and we wish to have the probability of the estimation error being greater than 0.1 units be equal to 0.1. What sample size is required if the standard deviation $\sigma = SD(X_i)$ is (i) $\sigma = 0.5$ or (ii) $\sigma = 0.2$? You can treat $\bar{X}$ as being normally distributed.

**Solution:** For this problem, I was thinking of the error as being the absolute difference between the estimate and the true value, $|\bar{X} - \mu|$. But if you use the signed error $\bar{X} - \mu$, that’s OK this time.

The following calculation gives the probability of the absolute error being greater than 0.1 units:

\[
P(|\bar{X} - \mu| > 0.1) = 2P(\bar{X} - \mu > 0.1)
\]
\[
= 2P(\sqrt{n}\frac{\bar{X} - \mu}{\sigma} > 0.1\sqrt{n}/\sigma)
\]
\[
= 2P(Z > 0.1\sqrt{n}/\sigma)
\]

We want an error larger than 0.1 to occur only 0.1 fraction of the time, so we set

\[
2P(Z > 0.1\sqrt{n}/\sigma) = 0.1,
\]

or

\[
P(Z > 0.1\sqrt{n}/\sigma) = 0.05.
\]

From a normal probability table, you can find that 1.64 is the point such that

\[
P(Z > 1.64) = 0.05.
\]

Thus we set $0.1\sqrt{n}/\sigma = 1.64$, or

\[
n = 16.4^2\sigma^2.
\]

For (i), we get $n = 67$, for (ii) we get $n = 11$.

When the variability in the data is lower, we don’t need as large of a sample size to control the errors.
2. Suppose we conduct a survey of peoples’ attitudes toward a certain product, obtaining independent data $X_1, \ldots, X_{100}$. We decide (somewhat arbitrarily) to use $\bar{X} \pm \sigma/2$ as a confidence interval for $\mu = E(X_i)$, where $\sigma = SD(X_i)$ (all the $X_i$ are assumed to have the same standard deviation). What is the coverage probability of this interval? You can treat $\sigma$ as being known. Hint: start with the definition of coverage probability and work backwards. You can treat $\bar{X}$ as having a normal distribution. If you need a normal distribution table, you can find one on the web.

Solution:

$$P(\bar{X} - \sigma/2 \leq \mu \leq \bar{X} + \sigma/2) = P(-\sigma/2 \leq \mu - \bar{X} \leq \sigma/2)$$

$$= P(-\sigma/2 \leq \bar{X} - \mu \leq \sigma/2)$$

$$= P(-\sqrt{n}/2 \leq \sqrt{n}(\bar{X} - \mu)/\sigma \leq \sqrt{n}/2)$$

$$= P(-5 < Z < 5)$$

The answer is around 0.9999994, but the real point is just that this interval is so wide that it nearly always covers the true value.

3. Suppose we are interested in the amount of carbon dioxide emitted from a certain type of vehicle. Since the vehicles of this type vary based on their age, maintenance history, driving history, etc., we can view the CO$_2$ emissions as a random variable $X$ with mean $\mu$ (our parameter of interest) and variance $\sigma^2$. An additional source of variance results from errors in measuring the CO$_2$ levels. We have two alternatives. A measurement $Y$ can be made that is highly accurate, essentially giving the actual value $X$, or a cheaper measurement $Z$ can be made that has variance $\sigma^2 + \tau^2$. Suppose the accurate measurements cost $10 dollars per vehicle, and the less accurate measurements cost $5 per vehicle. If we have a fixed amount of money to spend for this project, under what conditions on $\tau^2$ should we use the cheaper measurements?

Solution: For any fixed budget, we will be able to have two times greater sample size if we use the cheaper measurement. We’ll write the sample sizes as $n$ and $2n$, and the corresponding variances are $\sigma^2$ and $\sigma^2 + \tau^2$. Thus the variances of the sample mean (which we use to estimate the population mean) are $\sigma^2/n$ and $(\sigma^2 + \tau^2)/(2n)$. So, we will use the cheaper measurements when

$$\frac{\sigma^2}{n} > \frac{(\sigma^2 + \tau^2)}{(2n)},$$

which is equivalent to $\tau^2 < \sigma^2$.

4. Suppose we are comparing two medical procedures in a study in which half of the subjects are treated with one procedure and half are treated with the other procedure. The total sample size is 100, and both treatments have a population standard deviation
of 1 unit. Using a two-sample Z-test (treating the variances as known), what value of $D$ (the difference between the group-level means) would result in a p-value of 0.05?

**Solution:** The denominator of the test statistic $T$ is

$$\sqrt{\frac{1}{50} + \frac{1}{50}} = \frac{1}{5}.$$

A p-value of 0.05 results when the test statistic equals 2. Solving $5D = 2$ yields $D = 2/5$. 