Worksheet 01: Z-Test

The worksheets in this class are designed to guide you through developing the core mathematical results of the material. These include both important results and space for you to work out intermediate results. For much of the semester we will be deriving a very specific set of results on which classical statistics is based. This is a somewhat different approach to other mathematics courses you may have taken where you learn general techniques that can be applied to a seamingly endless sequence of similar exercises.

Today, we will start by definining some of the core terminology that extends probability theory to statistical theory. In statistics, we are typically describing a repeated measurement of a random process. The goal is to use the observations of those measurements to better understand the process. As a way to model this process, we will characterize the repeated measurements a set of independent and identically distributed random variables from some distribution $\mathcal{G}$. In symbols, we have $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{G}$. We call this a **random sample** of size n. Each component X_i is called an **observation**.

Often, we will be concerned with the expected value and variance of the distribution. These are sometimes known as the **population mean** and **population variance**. Typically, we will use μ_X and σ_X^2 to stand for these quantities. We can drop the X subscript if it is clear which distribution we are working with. The **sample mean** is a function of the random sample. It is denoted and defined by:

$$\bar{X} = \frac{1}{n} \times [X_1 + \cdots + X_n] = \frac{1}{n} \times \sum_{i=1}^{n} X_i$$

Similarly, the **sample variance** is given by:

$$S_X^2 = \frac{1}{n-1} \times \sum_{i=1}^{n} \left[X_i - \bar{X}\right]^2$$

Notice that both the sample mean and the sample variance are themselves random variables. Today, we will focus on understanding the sample mean.

Specifically, we want to consider the common task of estimating the value of μ_X. That is, we observe a set of observations of some random process and want to estimate the average value of the random process based on the sample average of the data that we collected. We will work through some theoretical results before applying this to a specific dataset.

Q01. What is the expected value of the sample mean? Write this in terms of μ_X.

$$\begin{aligned}\mathbb{E}\bar{X} &= \mathbb{E}\left[\frac{1}{n} \times \sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \times \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \times \left[\sum_{i=1}^{n} \mathbb{E}X_i\right] \\ &= \frac{1}{n} \times \left[\sum_{i=1}^{n} \mu_X\right] \\ &= \frac{1}{n} \times n \cdot \mu_X = \mu_X.\end{aligned}$$

Q02. What is the variance of the sample mean? Write this in terms of σ_X^2.

$$\begin{aligned}\mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\frac{1}{n} \times \sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2} \times \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2} \times \left[\sum_{i=1}^{n} \mathrm{Var} X_i\right] \\ &= \frac{1}{n^2} \times \left[\sum_{i=1}^{n} \sigma_X^2\right] \\ &= \frac{1}{n^2} \times n \cdot \sigma_X^2 = \frac{\sigma_X^2}{n}.\end{aligned}$$

For a moment, assume that $\mathcal{G}$ is a normal distibution. **Q03.** What is the distribution of $\bar{X}$? An independent sum of normal distributions is still normal, as is the linear scaling of a normal distribution. Therefore, $\bar{X}$ should also be normal. From the previous two results, we have that $\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$, since those are its mean and standard deviation.

Q04. Through what results does the previous result hold regardless of the specific distribution of $\mathcal{G}$ for a large enough n? The **central limit theorem** tells us that we can approximate $\bar{X}$ by a normal distribution as long as μ_X and σ_X^2 are finite and n is sufficently large.

We typically use the letter Z to indicate a random variable from a standard normal distribution. That is, $Z \sim N(0, 1)$. Notice that we can always rescale a normal distribution to be a standard normal. For example, if $Y \sim N(a, b^2)$, then:

$$Z = \frac{Y - a}{b} \sim N(0, 1).$$

You can prove this by first computing the expected value and variance of the new quantity and then justifying that this scaling must still be normal.

Q05. Rewrite the sample mean as a random variable Z that has a standard normal distribution.

$$Z = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}.$$

There is no closed form version of the cdf of a standard normal in terms of simple function. We can approximate specific cut-off values though using numerical techniques (we will see how to do this directly in R later). For example, we have the following bound on Z being a distance of 2.58 or greater away from the origin:

$$\mathbb{P}\left[|Z| > 2.58\right] \approx 0.01$$

It will often be helpful to have a short symbol for the (inverse) cdf of the standard normal, which we define using the symbol z_α by:

$$\mathbb{P}\left[Z > z_\alpha\right] = \alpha.$$

From the symmetry of the normal, we have:

$$\mathbb{P}\left[|Z| > z_{\alpha/2}\right] = \mathbb{P}\left[|Z| > z_{1-\alpha/2}\right] = \alpha.$$

And we see that $z_{1-0.01/2} = 2.58$.

Q06. Using the previous results, produce a bound showing how close you expect $\bar{X}$ to be away from μ_X 99% of the time that you sample n observations from the distribution $\mathcal{G}$.

$$\begin{aligned} \mathbb{P}\left[\left|\frac{\bar{X} - \mu_X}{\sqrt{\sigma_X^2 / n}}\right| > 2.58\right] &\approx 0.01 \\ \mathbb{P}\left[|\bar{X} - \mu_X| > 2.58 \cdot \sqrt{\sigma_X^2 / n}\right] &\approx 0.01 \end{aligned}$$

So far, this has all been very abstract. Let's consider a concrete example. There is a recent diet fad where people eat nothing but potatoes in order to lose weight. Consider observing 16 randomly choosen individuals that have attempted this diet. We have the following (sorted for convenience) amounts of weight lost in kilograms over a 45-day period. Note that a negative score means that someone gained weight.

```
-4.6, -0.5, -0.5, 0.1, 0.1, 1.1, 1.9, 2.6, 3,
      3.2, 3.5, 3.7, 4.2, 5.4, 6.5, 6.8
```

We can compute the sample mean and sample variance as follows (note that we typically convert the upper-case letters used in the theoretical results to lower-case letters when describing observed quantities):

$$\bar{x} = 2.28, \quad s_X^2 = 8.7.$$

Q07. Based on the data, what would be your best guess as to the amount of weight that is lost of average with the potato diet? Why is this guess justified by the theoretical results? The best guess based on our current work would be to guess that the population mean of the underlying distribution of weight loss in general is equal to the sample mean of 2.28. This is justified because we know that the sample mean will be equal to the population mean on average and that the variance will limit to zero as the sample size increases to infinity.

We can do a lot more than just justify the best guess of the unknown amount of weight lost with the potato diet. In general, a $(1-\alpha)\%$ **confidence interval** for an unknown quantity γ is a pair of random variables L_γ and U_γ such that:

$$\mathbb{P}[L_\gamma \leq \gamma \leq U_\gamma] \geq 1 - \alpha.$$

Q08. From the previous results, what random variables could we use to create a 99%-confidence interval for the mean μ_X?

$$L_\mu = \bar{X} - 2.58 \cdot \sqrt{\sigma_X^2 / n}$$
$$U_\mu = \bar{X} + 2.58 \cdot \sqrt{\sigma_X^2 / n}.$$

One problem with using the confidence interval you derived above is that it requires already knowing the variance of $\mathcal{G}$. We will see the correct way to deal with this next time. For now, assume that we know the variance of the potato data is 3^2. **Q09.** Construct a 99%-confidence interval for the average amount of weight lost using the potato diet from our data. We have:

$$L_\mu = 2.28 - 2.58 \cdot \sqrt{3^2/16} = 0.345$$
$$U_\mu = 2.28 + 2.58 \cdot \sqrt{3^2/16} = 4.215.$$

Q10. What random variables could we use to create an $(1-\alpha)\%$-confidence interval for the mean μ_X in terms of $z_{1-\alpha/2}$?

$$L_\mu = \bar{X} - z_{1-\alpha/2} \cdot \sqrt{\sigma_X^2 / n}$$
$$U_\mu = \bar{X} + z_{1-\alpha/2} \cdot \sqrt{\sigma_X^2 / n}.$$