5 Probability, Sampling Distributions, Central Limit Theorem

As with last week, most of this material is covered quite nicely in SW, so I plan to stick with the text very closely. We will do a quite a bit of computer work to accompany this material, both during lecture and the lab.

Random Variables

SW Section 3.7.

This is our model for sampling from a population. If we sample (from either a categorical or quantitative) population, we write Y = the value obtained. If we sample n values from a population, the values obtained are $Y_1, Y_2, \ldots, Y_n$. The population we sampled from has a mean μ and standard deviation σ (we'll force even categorical variables into such a structure) – those are the mean and standard deviation of the random variable as well. Don't worry about the more mathematical treatment in SW.

Binomial Distribution

SW Section 3.8.

Eric covered the Independent-Trials Model with you in lab. The binomial distribution lays out probabilities for all the possible numbers of successes in n independent trials with probability p of Success each trial. This is a new population with a mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. The model is important, but don't worry about all the formulae in SW. Minitab does a great job of calculating probabilities when needed.

Normal Distribution

SW Sections 4.1-3.

Eric also covered this in lab. We want to revisit Figure 4.7, the standard normal Z, and the Standardization Formula $Z = \frac{Y-\mu}{\sigma}$ on p. 124. The figures on p. 125 are a valuable working guide. We will work a couple of examples, including Minitab calculations.

Normal distributions pop up in many more situations than you would expect. We need to be able to use them.

Assessing Normality

SW Section 4.4.

Eric will cover this in the lab. The normal probability plot is a widely used tool. SW do not talk about box plots here, but those also serve as valuable tools. We will be making the assumption many times that we sampled from a normal population. The assumption really does matter, so we need methods to assess it.

Sampling Distribution of $\overline{Y}$

SW Section 5.3.

If we randomly sample (SRS) n values $Y_1, Y_2, \ldots, Y_n$ from a quantitative population and calculate $\overline{Y}$, then $\overline{Y}$ depends upon the random sample we drew — if we drew another random sample

we would get a different value of $\overline{Y}$. This means $\overline{Y}$ is a random variable, i.e. it is a single random number sampled from *some* population.

From what population is $\overline{Y}$ drawn? It most certainly is not the same as the population $Y_1, Y_2, \ldots, Y_n$ come from (the possible values may not even be the same). It is a new population called the ***sampling distribution*** of $\overline{Y}$. I'll spare you any derivations and just cite some results.

Mean and Standard Deviation of $\overline{Y}$

If the population $Y_1, Y_2, \ldots, Y_n$ are sampled from has mean μ and standard deviation σ, the sampling distribution of $\overline{Y}$ has mean $\mu_{\overline{Y}} = \mu$ and standard deviation $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$. On average $\overline{Y}$ values come out the same place (μ) as the Y_i values, but they tend to be closer to μ than are the individual Y_i, since the standard deviation is smaller.

Shape of Sampling Distribution of $\overline{Y}$

This is the part with a lot of mathematics behind it. There are two cases when we can treat $\overline{Y}$ as if it was sampled from a Normal Distribution (and we know how to use normal distributions!):

1. If the population $Y_1, Y_2, \ldots, Y_n$ were sampled from is normal, no matter how small n is,
2. if n is large, almost no matter what the shape of the population from which $Y_1, Y_2, \ldots, Y_n$ were sampled.

We cannot say what the shape is for small n unless we originally sampled from a normal distribution. This is why we worry so much about assessing normality with boxplots and normal probability plots. Part 2 is the **Central Limit Theorem**.

Let us go over Examples 5.9 and 5.10. We will do a few simulations in Minitab to demonstrate the preceding results.

Sampling Distribution of $\hat{p}$

SW Sections 5.2, 5.5

This is how we really use the Independent-Trials Model, and the way we think of binary response variables. We now randomly sample n individuals from a population where every value is either a S or F (just generic labels). The proportion of S's in the population is p, the proportion of S's in the sample is $\hat{p}$. Again, $\hat{p}$ is a random variable since it depends upon the random sample, so it has a sampling distribution. What population is $\hat{p}$ sampled from?

The amazing result is that if n is large, we can assume $\hat{p}$ was drawn from a *Normal* population with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. For this to hold we need $np \geq 5$ and $n(1-p) \geq 5$.

We will use Minitab to demonstrate this, and do a few calculations.