

The Distribution of a Statistic


Let's forget science for the moment and consider the simple experiment of flipping a single coin (fair coin, fair flip), with the result hidden from you. Most folks, if asked, would say that there is a 50% chance that the coin is heads. What is the basis for that line of thinking? (There are subtle, perhaps only semantic, questions regarding the use of the word "chance" once the coin is flipped, but we won't go there just now.) The basis for claiming that there is a 50% chance that the coin is heads is simply the understanding that in a very large number of flips of a fair coin, heads will result about 50% of the time. We intuitively (and often without explicitly realizing it) apply our understanding of the long-run of coin flips to the current one.

As a basis for making a judgment regarding the current situation, one imagines what the outcomes would be if the experiment were repeated a large number of times. Note that one doesn't need to actually do the large number of repeats of the experiment to use it as a basis for inference.

Back to science. For convenience, I will stick with the example of the sample mean of a random sample of observations. Here, the sample



mean is the chosen statistic; the population mean is the parameter being estimated. The distribution of the mean is the distribution of values of that statistic you would get by repeating the study a huge number of times. Many (maybe most) statistics authors call this the “sampling distribution.” I find the adjective “sampling” adds more confusion than clarity, and find that omitting it doesn’t cause any problems at all (other than the need to explain why I don’t use the term). In practice, this idea of repeating a study a large number of times is absurd: it’s simply not possible to repeat studies under identical conditions a large number of times. Nonetheless, just as in the coin flipping example, we can conceptualize the repeats.

In principle, if one repeated the study an infinite number of times, the average of all the sample means would in fact be the true population mean. This is true because the procedure (simple random sample from some population, sample mean as chosen statistic) is known to be unbiased. (In fact, the previous two sentences are just different ways of saying the same thing).

In fact, this notion of the long-term predictability (of certain features) of random events underpins all of statistical inference  .

Assuming Normality

For the most commonly used statistics (means, differences in means, proportions, differences in proportions, regression slopes), we often choose to assume that a statistic has an approximately Normal distribution. Indeed, most of our regularly used statistical inference tools depend on this assumption. This assumption will be valid if the underlying randomness in the data is approximately Normal or if our sample size is large enough (how large is large enough depends on the underlying distribution).


Often biologists encounter data that are skewed; count data are a common example  . Another commonly encountered type of (way seriously!) non-Normal data is Binomial data. You have Binomial data if your response is recorded as a dichotomy: success/failure, male/female, presence/absence, etcetera  .

The SD of a Statistic

Critical to statistical inference is the ability to estimate the SD of your statistic. This SD is the variation among values of your chosen statistic in a large number (in principle, infinite!) of repeats of your study.

Let's assume for now that you have reason to believe that your statistic has (at least approximately) a Normal distribution. The value of your statistic itself stands as an estimate of the mean of that distribution. How can we possibly estimate the SD of that distribution when we only have a single value (your actual mean, or whatever) from it?

It turns out that the variation in the distribution of a given statistic is related to the variation in the raw data (in various ways: the formulae are different depending on the choice of statistic and so on). For example, the formula to estimate the SD of a mean from a random sample is just $SD(\text{mean}) = \frac{SD}{\sqrt{n}}$, where SD (with no qualifying parenthetical phrase) is the standard deviation among the sample values, and n is the

sample size.  .

A common name for what I call the SD of a statistic is the Standard Error (SE) of a statistic. I don't like the term (there is no "error" involved, other than random "error", which is present everywhere, so why bother highlighting it here?). I find that when I use SD for the standard deviation of one distribution and SE for the standard deviation of another, people get confused. Down with confusion! Unfortunately, I doubt my crusade will get very far...

