# Data Science Career Track

## *The Art of Statistics*, Chapter 9: Putting probability and statistics together

## Take-Away Notes

---

We've already seen the idea of a **random variable:** a single data-point taken from a probability distribution that's describable with parameters. We typically have a mass of data, that's characterized with the summary statistics of (for example) means and medians. In this chapter, we treat *those very statistics* as random variables, drawn from their own distributions.

- We can use probability theory to determine the **sampling distribution of summary statistics**, before using these to derive the formulae for confidence intervals.
- The mean of a random variable is known as its **expectation**. The standard deviation of a statistic is termed the **standard error**, to distinguish it from the standard deviation of the population from which it derives.
- If there are *n* independent possibilities in which an event can occur, each with the same probability, the observed number of events has a **binomial distribution**.

**Funnel plots** are a very useful method for plotting a set of observations from different units against a measure of their precision, where units could be things like institutions, areas, or studies.

- Often, two funnels show where we'd expect 95% and 99.8% (respectively) of observations to lie, on the assumption that there really are no underlying differences between the units.
- If the distribution of the observations is roughly normal, the 95% and 99.8% **control limits** are just the mean +- two and three standard errors, respectively.

This document is authorized for use only by Shirley Li (sl496@duke.edu). Copying or posting is an infringement of copyright.

The **Law of Large Numbers** is the claim that the sample mean of a set of random variables tends towards the population mean. If, for example, you keep on flipping a balanced coin, the proportion of each outcome will get closer and closer to 50% heads and 50% tails; the observed proportion *tends to* the true underlying chance of a head. The **Central Limit Theorem** entails that sample means and other summary statistics can be assumed to have a normal distribution for large samples.

**Please note:**
1. Systematic error from non-random causes is *not incorporated* by margins of error, and external judgment is required to assess these.
2. Even if we've observed *all the data*, we can still usefully calculate confidence intervals, which in this situation represent *uncertainty about the parameters of an underlying metaphorical population.*

This document is authorized for use only by Shirley Li (sl496@duke.edu). Copying or posting is an infringement of copyright.