7.1 Statistical Inference

Probability vs Mathematical Statistics

The Theory of Probability is the foundation for Mathematical Statistics, but the former is not simply a subset of the latter; the two subjects exhibit important philosophical differences.

- While Probability is inspired experience, it is fundamentally rooted in **axiom and theory**. In probability, we might posit a model and study its properties, but rarely ask whether it fits empirical observation.
- Conversely, Mathematical Statistics is chiefly concerned with data, using the theory of probability to formulate plausible models based on observations.
- Probability begins with a sample space and experiment, and investigates the liklihood of possible outcomes. *Ex: How likely is it that it takes at least 10 rolls to produce a 1 on a fair 6-sided die.*
- Statistics begins with outcomes, and investigates the possible frameworks that could have produced that data. *Ex: Given that a 6-sided die did not roll a 1 until the 10th roll, how likely is it that the die is fair?*
- In Probability, no model is wrong (by assumption).
- In Statistics, every model is wrong (but some are useful).

Ex 1: Consider a box filled with 100 tickets, some of which are red and some of which are blue.

In probability, we might *assume* that a proportion p of the tickets are red, and investigate the variable X counting the number of red tickets, if we draw n tickets one-by-one with replacement. Recall that this is exactly the story of the Binomial distribution, so $X \sim \text{Bin}(n, p)$ and

$$P(X = k) = \binom{n}{k} p^k (1-p)^{t-k}.$$

We might wonder what the shape of this distribution looks like as $n \to \infty$. By the Central Limit Theorem, X is approximately Normal, with mean np and variance $np(1-p)$.

In statistics, instead we might *observe* that when we sampled n times with replacement, exactly k of the tickets were red. Before sampling, we make no assumptions about the true proportion of red tickets, and try to infer properties about the unknown value p.

Let's see this in action.

Ex 2: Suppose each red ticket in the box is worth \$1 and each blue ticket is worth \$5. I will sell you the box for \$200. This would correspond to having 75 red and 25 blue tickets.

I won't tell you exactly how many red tickets are in the box, but I will let you peak by drawing 8 tickets from the box. You can even decide whether you want to draw them with, or without replacement.

- In groups, discuss benefits/drawbacks of drawing with and without replacement. Which is preferable? Why? *Hint: Perhaps consider the extreme cases when $n = 1$ and $n = 100$.*

Now, let's actually sample! [Do sampling].

In groups, discuss the following:

- What is your best estimate for the proportion of red tickets?
- Why is this your best estimate?

- How certain are you that this is exactly the true proportion?
- How certain are you that this is close to the true proportion?
- Quantify your certainty by using the 68-95-99.7 rule, along with the Central Limit Theorem (assume that 8 is "big enough" to use CLT here).
- Suppose there were actually 75 red tickets. How likely would it be to see a result exactly like the one you did?
- How likely would it be to see a result like the one you did, or more extreme?
- Will you buy the box?

Models and Parameters and Statistics

Data comes in many forms. Often, it consists of a list of numeric or character values. But it can also be organized more rigidly, as a matrix or array. Or it can present itself non-numerically, as a function, graph, or image.

A model is a structure for data production:

Def: We are given a random experiment with sample space Ω and a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ whose components take values in a space D (often $D = \mathbb{R}$). For an outcome $\omega \in \Omega$, we refer to $\mathbf{X}(\omega)$ as the **observations** or **data**. A **model** $\mathcal{P}$ is a proposed family of joint distributions on D^n to which $\mathbf{X}$ could belong.

Ex 3: In the preceding example discussing red and blue tickets, we define a sequence of random variables $X_1, X_2, X_3, \ldots$ where X_n counts the number of defects in a sample of n items. The family of possible joint distributions of this sequence is indexed by the parameter p; for fixed p, the variables are independent, with $X_n \sim \text{Bin}(n, p)$. This is also the conditional distribution of $X_1, X_2, \ldots$ given p.

Def: **Statistical Inference** is a procedure that produces a probabilistic statement about some or all parts of a statistical model.

Ex 4: In the preceding example, we might make statistical inference by producing a variable Y which is a function $X_1, \ldots, X_m$ so that $P(Y \leq p \,|\, p) = 0.99$, or which takes the value θ on average, or which has .95 probability of being no more than a distance of 0.1 from θ.

Much of statistical inference and model building involves the specification of parameters, which is formalized as...

Def: A **parameter** is a characteristic or combination of characteristics whose values determine the joint distribution of the variables in the model. The set Ω of all possible values for the parameter(s) is called the parameter space.

Ex 5: In the previous example, the proportion p of red tickets in the box is a parameter, since it determines the distribution of X_n, the number of red tickets in a sample of size n. We know that the number of red tickets must be an integer between 0 and 100, so the parameter space for p is

$$\Omega = \left\{0, \frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}, 1\right\}.$$

Statisticians create models and parameterizations. But the values of parameters will almost always be unknowable. The aim of Statistics is to use data inductively to narrow down the value of the parameter.

Def: A **statistic** T is a function from a random vector $\mathbf{X}$ on a sample space to a collection of values $\mathcal{T}$ (usually a subset of n-dimensional space, and often just $\mathbb{R}$).

Ex 6: In the ticket box example, both the number of red tickets X and the fraction of red tickets $T(X) = \frac{X}{n}$ are statistics. So is the random variable Y which takes the value 1 if the sample contains at least 1 red ticket, and 0 otherwise.